THE OPTIMISED SCHWARZ METHOD AND THE TWO-LAGRANGE MULTIPLIER METHOD FOR HETEROGENEOUS PROBLEMS

by

Neil Greer

Submitted for the degree of
Doctor of Philosophy

Department of Mathematics

School of Mathematical and Computer Sciences
Heriot-Watt University

April 2017

Abstract

In modern science and engineering there exist many heterogeneous problems, in which the material under consideration has non-uniform properties. For example when considering seepage under a dam, water will flow at vastly different rates through sand and stone. Mathematically this can be represented as an elliptic boundary value problem that has a large jump in coefficients between subdomains. The optimised Schwarz method and the related two-Lagrange multiplier method are non-overlapping domain decomposition methods that can be used to numerically solve such boundary value problems.

These methods work by solving local Robin problems on each subdomain in parallel, which then piece together to give an approximate solution to the global boundary value problem. It is known that with a careful choice of Robin parameter the convergence of these methods can be sped up.

In this thesis we first review the known results for the optimised Schwarz method, deriving optimised Robin parameters and studying the asymptotic performance of the method as the mesh parameter of the discretisation is refined and the jump in coefficients becomes large.

Next we formulate the two-Lagrange multiplier method for a model two subdomain problem and show its equivalence to the optimised Schwarz method under suitable conditions. The two-Lagrange multiplier method results in a non-symmetric linear system which is usually solved with a Krylov subspace method such as GMRES. The convergence of the GMRES method can be estimated by constructing a conformal map from the exterior of the field of values of the system matrix to the interior of the unit disc.

We approximate the field of values of the two-Lagrange multiplier system matrix by a rectangle and calculate optimised Robin parameters that ensure the rectangle is "well conditioned" in the sense that GMRES converges quickly. We derive convergence estimates for GMRES and consider the behaviour asymptotically as the mesh size is refined and the jump in coefficients becomes large.

The final part of the thesis is concerned with the case of heterogeneous problems with many subdomains and cross points, where three or more subdomains coincide. We formulate the two-Lagrange multiplier method for such problems and consider known preconditioners that are needed to improve convergence as the number of subdomains increases.

Throughout the thesis numerical experiments are performed to verify the theoretical results.

Acknowledgements

I would first like to thank my supervisor Dr Sébastien Loisel for the indispensable help and guidance he has provided. I also wish to thank my family for their support and encouragement. Finally I would like to thank the Engineering and Physical Sciences Research Council (EPSRC) who provided the funding for my PhD studies.

ACADEMIC REGISTRY

Research Thesis Submission

Name*:*			
School:			
Version: *(i.e. First, Resubmission, Final)*		Degree Sought:	

Declaration

In accordance with the appropriate regulations I hereby submit my thesis and I declare that:

1) the thesis embodies the results of my own work and has been composed by myself
2) where appropriate, I have made acknowledgement of the work of others and have made reference to work carried out in collaboration with other persons
3) the thesis is the correct version of the thesis for submission and is the same version as any electronic versions submitted*.
4) my thesis for the award referred to, deposited in the Heriot-Watt University Library, should be made available for loan or photocopying and be available via the Institutional Repository, subject to such conditions as the Librarian may require
5) I understand that as a student of the University I am required to abide by the Regulations of the University and to conform to its discipline.
6) I confirm that the thesis has been verified against plagiarism via an approved plagiarism detection application e.g. Turnitin.

* *Please note that it is the responsibility of the candidate to ensure that the correct version of the thesis is submitted.*

Signature of Candidate*:*		Date:	

Submission

Submitted By *(name in capitals):*	
Signature of Individual Submitting:	
Date Submitted:	

For Completion in the Student Service Centre (SSC)

Received in the SSC by *(name in capitals):*			
Method of Submission *(Handed in to SSC; posted through internal/external mail):*			
*E-thesis Submitted (**mandatory for final theses**)*			
Signature:		Date:	

Please note this form should be bound into the submitted thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Problems arising in physics and engineering often require the solution of elliptic boundary value problems (BVPs), which we wish to solve numerically. Once the BVP has been discretised with a suitable method, such as finite difference, finite element or finite volume, we need to solve a linear system of the form:

$$A\boldsymbol{u} = \boldsymbol{f}, \tag{1.1}$$

where A is a sparse, symmetric positive definite matrix. To achieve a desired level of accuracy the discretisation of the BVP will need to be very fine and as a result the matrix A will be very large. To solve system (1.1) with a direct method would be costly and impractical due to fill in so instead we use an iterative solver.

Domain decomposition methods (DDMs) are a group of iterative methods that partition the physical domain of the BVP into smaller subdomain BVPs which are solved and pieced together to construct a global solution. Moreover by taking advantage of modern computer architecture where machines can have thousands of cores, DDMs allow the parallel solution of the subdomain problems by assigning each one to a separate processor.

The earliest DDM due to Hermann Schwarz in 1870 was introduced as a proof technique

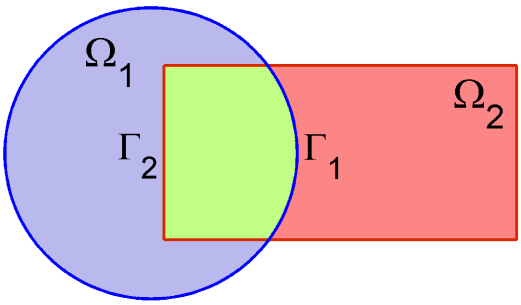

Figure 1.1: domain decomposed into two overlapping subdomains

for the Dirichlet principle. The Dirichlet principle states that the solution u of the Laplace equation $\Delta u = 0$ with Dirichlet boundary conditions $u = g$ on $\partial\Omega$ minimises the energy functional $\int_\Omega |\nabla u|^2 dx$. At the time the Dirichlet principle could be proved on simple domains such as circles and rectangles by using Fourier analysis. In [60] Schwarz proved that the Dirichlet principle held on more general domains, constructed from the overlapping union of elementary subdomains like circles and rectangles. An example general domain is shown in Figure 1.1. By imposing Dirichlet boundary conditions on the interfaces, $\Gamma_i = \partial\Omega_{3-i} \cap \Omega_i$ for $i = 1, 2$, Schwarz proposed the *alternating Schwarz method*:

$$\begin{cases} \Delta u_1^{n+1} = 0 \quad in \quad \Omega_1 \\ u_1^{n+1} = 0 \quad \text{on} \quad \partial\Omega_1 \cap \partial\Omega \\ u_1^{n+1} = u_2^n \quad \text{on} \quad \Gamma_1 \end{cases}, \quad \begin{cases} \Delta u_2^{n+1} = 0 \quad in \quad \Omega_2 \\ u_2^{n+1} = 0 \quad \text{on} \quad \partial\Omega_2 \cap \partial\Omega \\ u_2^{n+1} = u_1^{n+1} \quad \text{on} \quad \Gamma_2 \end{cases}$$

As these subproblems are just BVPs on a circle and a rectangle they could be solved using known Fourier series methods. Schwarz proved convergence in that $\lim_{n\to\infty} u_i^n = u|_{\Omega_i}$.

Over 100 years later in the 1980s with the advent of multi-core computers interest in Schwarz's method was renewed with the observation that it could be implemented in

parallel with the modification:

$$\begin{cases} \Delta u_1^{n+1} = 0 \quad in \quad \Omega_1 \\ u_1^{n+1} = 0 \quad \text{on} \quad \partial\Omega_1 \cap \partial\Omega \\ u_1^{n+1} = u_2^n \quad \text{on} \quad \Gamma_1 \end{cases}, \quad \begin{cases} \Delta u_2^{n+1} = 0 \quad in \quad \Omega_2 \\ u_2^{n+1} = 0 \quad \text{on} \quad \partial\Omega_2 \cap \partial\Omega \\ u_2^{n+1} = u_1^n \quad \text{on} \quad \Gamma_2 \end{cases}$$

This method is known as the *parallel Schwarz method* and together with the alternating variant is referred to as a *classical Schwarz method* (CSM).

However several drawbacks of the CSMs were observed. Firstly, CSMs require that the subdomains overlap to guarantee convergence. In practice problems arise with naturally non-overlapping subdomains. Consider the model problem

$$\begin{cases} -\nabla \cdot (a(\boldsymbol{x})\nabla u) = f \qquad \text{in} \qquad \Omega \\ \qquad\qquad\qquad u = 0 \qquad \text{on} \qquad \partial\Omega. \end{cases} \tag{1.2}$$

The open set $\Omega \subset \mathbb{R}^n$ is partitioned into two non-overlapping Lipschitz subdomains Ω_1, $\Omega_2 \subset \Omega$, such that

$$a(\boldsymbol{x}) = \begin{cases} \alpha_1 \alpha_0(\boldsymbol{x}) \quad \text{for} \quad \boldsymbol{x} \in \Omega_1 \\ \alpha_2 \alpha_0(\boldsymbol{x}) \quad \text{for} \quad \boldsymbol{x} \in \Omega_2, \end{cases}$$

where α_1, $\alpha_2 \in \mathbb{R}^+$ and $\alpha_0(\boldsymbol{x})$ is a continuous function with $0 < \alpha_{\min} < \alpha_0(\boldsymbol{x}) < \alpha_{\max} < \infty$. Here we assume that $\alpha_{\max}/\alpha_{\min} \ll \infty$, while α_1 and α_2 can be of a vastly different magnitude (e.g. $\alpha_{\max}/\alpha_{\min} < 10$ and $\alpha_1/\alpha_2 = 10^6$). We allow $a(\boldsymbol{x})$ to be otherwise arbitrary, but note that if $a(\boldsymbol{x})$ has jumps then the PDE operator $\nabla \cdot (a(\boldsymbol{x})\nabla u)$ does not have meaning in the strong sense but must be interpreted in the weak (variational) sense, as we will do shortly.

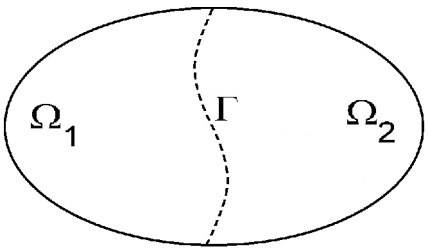

Figure 1.2: non-overlapping decomposition of a general domain into two general subdomains

If $f \in L^2(\Omega)$, then the weak formulation of (1.2) has a solution $u \in H_0^1(\Omega)$. If we define the interface between the subdomains by

$$\Gamma = \Omega \cap (\partial\Omega_1 \cup \partial\Omega_2),$$

then $\Omega = \Omega_1 \cup \Omega_2 \cup \Gamma$ and in this case $\Gamma_1 = \Gamma_2$. The requirement that Ω_i, for $i = 1, 2$, are Lipschitz ensures that Γ is also Lipschitz.

When $\alpha_1 \neq \alpha_2$ there is a discontinuity across Γ and (1.2) corresponds to a problem in heterogeneous media. The jump in coefficients of the problem naturally defines a non-overlapping decomposition of the physical domain Ω. Figure 1.2 shows an example of a non-overlapping decomposition of a general domain.

Other drawbacks of the CSMs include that convergence can be slow especially when the overlap is small and that they do not converge at all for certain problems such as the Helmholtz equation.

To remedy the problems of CSMs and obtain a DDM that can be implemented using non-overlapping subdomains it was observed that instead of imposing Dirichlet conditions across the interface, general linear operators $\mathcal{B}_i$ for $i = 1, 2$, can be imposed:

$$\begin{cases} \Delta u_1^{n+1} = 0 \quad in \quad \Omega_1 \\ \mathcal{B}_1(u_1^{n+1}) = \mathcal{B}_1(u_2^n) \quad \text{on} \quad \Gamma \end{cases}, \quad \begin{cases} \Delta u_2^{n+1} = 0 \quad in \quad \Omega_2 \\ \mathcal{B}_2(u_2^{n+1}) = \mathcal{B}_2(u_1^n) \quad \text{on} \quad \Gamma \end{cases}$$

The main idea behind DDMs and how the different methods arise is the choice of these linear operators and how to transfer information between the subdomains, such that the iterates applied to the subproblems piece together to give the global solution. Popular DDMs including Dirchlet-Neumann and Neumann-Neumann are constructed by choosing suitable forms for $\mathcal{B}_1$ and $\mathcal{B}_2$. We refer the reader to [50, 56, 63] for treatises on the subject of DDMs.

In [44] Lions proposed a non-overlapping variant of Schwarz's method that imposes Robin transmission conditions, $\mathcal{B}_i(u_i^n) = (\partial_{n_i} + p)u_i^n$, on Γ. Here ∂_{n_i} denotes the directional derivative with respect to the outward pointing normal n_i of $\partial\Omega_i$. Convergence was proved using energy estimates for any choice of Robin parameter $p > 0$. Though Lions didn't provide one he observed that a careful choice of parameter would lead to faster convergence.

Optimised Schwarz methods (OSM), [23, 26], aim to find such optimal Robin parameters to speed up convergence. See [21] for a full history of the various Schwarz methods. Convergence of the OSM is usually proved using Fourier analysis, [20], and so the subdomains considered are restricted to being rectangular. In [48] the author analyses the convergence for more general subdomains by considering the spectral radius of the interface operator that is expressed in terms of the Dirichlet to Neumann map (also known as the Poincaré-Steklov operator).

A DDM closely related to the OSM is the *two-Lagrange multiplier* (2LM) method introduced in [19]. The main idea behind the 2LM method is not to introduce an iteration explicitly for the subdomain solutions, but to replace the large linear system (1.1) with the smaller equivalent system

$$A_{2LM}\boldsymbol{\lambda} = \boldsymbol{c}, \tag{1.3}$$

where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers that are used to solve local Robin problems on the subdomains in parallel. The 2LM method is more suited than the OSM in dealing

with cross points, which occur where three or more subdomains coincide. In [10, 45] it was shown, for general subdomains with cross points, that if the Robin parameter on the interface is chosen to be of $O(h^{-1/2})$ the condition number of the 2LM method system matrix is of $O(h^{-1/2})$, where h is the finite element parameter. Efforts have been made, for example in [24], to estimate the convergence of OSM in the presence of cross points. In the absence of cross points, when the subdomains are arranged in strips, it is known, [57], that the OSM and 2LM method are equivalent when a Richardson iteration is applied to system (1.3). In the case of many subdomains the convergence of the OSM and the 2LM method will deteriorate as the number of subdomains increases. A preconditioner will be required to transfer information globally, in [39] and [47] the authors develop preconditioners for the 2LM method in the presence of cross points.

Many of the results described thus far have been for problems in homogeneous media, i.e. there is no jump in coefficients between the subdomains. The OSM in heterogeneous media has been studied using Fourier analysis on rectangular subdomains in [22, 49] and by estimating the spectral radius of the interface operator acting on general subdomains in [13]. Both approaches show that with a suitable choice of Robin parameters the speed of convergence of OSM is faster when the jump in coefficients becomes larger. In this thesis we are interested in the 2LM method for heterogeneous problems in a general domain with general subdomains.

Even though system (1.1) may be symmetric the related 2LM method system (1.3) will be non-symmetric. While smaller than the original system, it is still too costly to solve the 2LM system directly so we use an iterative Krylov subspace method for non-symmetric systems such as GMRES (*Generalised Minimal RESidual*), [59]. There are many ways to estimate the speed of convergence of the GMRES method when solving a linear system. One such estimate is obtained by calculating a conformal map from the exterior of the field of values of the system matrix to the interior of the unit disc, [8], where the field of values

is a compact convex subset of $\mathbb{C}$ that contains the spectrum of a given matrix. As with the OSM method a careful choice of Robin parameter in the 2LM method can lead to faster convergence. It is our goal in this thesis to find such optimised Robin parameters for the 2LM method applied to heterogeneous problems.

Our original contributions in the thesis are as follows. We introduce novel formulations of the 2LM method for heterogeneous problems involving two subdomains and multiple subdomains with cross points (Chapter 3 and Chapter 6 respectively). We derive optimised Robin parameters for the 2LM method and OSM using the approach of estimating the field of values of a matrix and present asymptotic convergence theory for their performance (Chapter 5). Original numerical experiments are performed to confirm our theoretical results (Chapter 3, 6). The results from Chapter 3 and 5 were included in the paper [33]. Chapters 2 and 4 are review only and do not contain any original work.

The thesis is organised as follows. In Chapter 2 we review the OSM for problems in heterogeneous media and quote some known results for its convergence. In Chapter 3 we derive the 2LM method for heterogeneous problems and show its equivalence to the OSM in the case of two subdomains. In Chapter 4 we outline the GMRES method and the ways in which its convergence can be approximated. In Chapter 5 we provide optimised Robin parameters in the case of a heterogeneous problem with a general domain and two general subdomains. These parameters are used to estimate the convergence rate of GMRES applied to the 2LM system and study the asymptotic behaviour as the finite element parameter h becomes small and the jump between coefficients α_1 and α_2 becomes large, confirming the results with numerical experiments. In Chapter 6 we consider the 2LM method in the case of many subdomains and cross points, modifying some known preconditioners for the 2LM method for our heterogeneous problem and testing their effectiveness numerically.

Chapter 2

The optimised Schwarz method

2.1 Formulation and convergence results

Consider the model heterogeneous problem (1.2). For $i = 1, 2$, let a_i denote the restriction of coefficient $a(\boldsymbol{x})$ to subdomain Ω_i and u_i the restriction of the solution u to subdomain Ω_i. Then given initial guess u_i^0 the continuous form of the OSM iteration for $n = 1, 2, \ldots$ is: solve for $i = 1, 2$

$$\begin{cases} -a_i \Delta u_i^n = f & \text{in} \quad \Omega_i \\ u_i^n = 0 & \text{on} \quad \partial\Omega_i \cap \partial\Omega \\ (p_i + a_i \partial_{N_i}) u_i^n = \lambda_i^{n-1} = (p_i + a_{3-i} \partial_{N_{3-i}}) u_{3-i}^{n-1} & \text{on} \quad \Gamma, \end{cases} \tag{2.1}$$

where ∂_{N_i} denotes the directional derivative with respect to the outward pointing normal N_i of $\partial\Omega_i$. Though in practice the Robin parameters $p_i \in (0, \infty)$ could vary along the interface here we only consider the case where they are constant on Γ.

The version of the OSM we have presented above can be implemented in parallel, as system (2.1) can be solved simultaneously for each subdomain with the only exchange of

information needed after each iteration to update the Robin relation in the third line. The difficulty arises in calculating the normal derivatives, however in [7] the author presents a formula for updating the Robin relation that avoids computing these derivatives.

To study the convergence of the OSM when we have a jump in coefficients between the subdomains we consider the problem with a global domain given by $\Omega = \mathbb{R}^2$ and subdomains given by the half planes $\Omega_1 = (-\infty, 0) \times \mathbb{R}$ and $\Omega_2 = (0, \infty) \times \mathbb{R}$. Then the interface Γ is the y-axis and we let

$$a(\boldsymbol{x}) = \begin{cases} \alpha_1 & \text{for} \quad x < 0 \\ \\ \alpha_2 & \text{for} \quad x > 0 \end{cases}$$

here $\alpha_1 > \alpha_2$ and we require that the solution decays to zero at infinity.

Applying the OSM iteration given by (2.1) to this problem we have

$$\begin{cases} -\alpha_1 \Delta u_1^{n+1} = f \quad \text{in} \quad \Omega_1 \\ \\ (p_1 + \alpha_1 \partial_x) u_1^{n+1}(0, y) = (p_1 + \alpha_2 \partial_x) u_2^n(0, y) \quad \text{for} \quad y \in \mathbb{R} \end{cases} \tag{2.2}$$

$$\begin{cases} -\alpha_2 \Delta u_2^{n+1} = f \quad \text{in} \quad \Omega_2 \\ \\ (p_2 - \alpha_2 \partial_x) u_2^{n+1}(0, y) = (p_2 - \alpha_1 \partial_x) u_1^n(0, y) \quad \text{for} \quad y \in \mathbb{R} \end{cases} \tag{2.3}$$

where each subdomain solution decays to zero at infinity. Note that the normal derivatives here are just the derivatives in the positive and negative x direction.

By linearity we need only consider the case when $f = 0$. To see this let

$$e_i^n = u_i - u_i^n \qquad \text{for} \qquad i = 1, 2,$$

where u_i is the exact subdomain solution in Ω_i and u_i^n the approximate solution at step n. Then e_i^n is the error in Ω_i at step n and we have

$$-\alpha_i \Delta e_i^n = -\alpha_i \Delta u_i + \alpha_i \Delta u_i^n = f - f = 0 \qquad \text{in} \qquad \Omega_i$$

Hence we can view the subdomain solutions u_i^n in (2.2) and (2.3), with $f = 0$, as error terms.

Using the following *Fourier transform* in variable y:

$$\hat{u}(x,k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} u(x,y) e^{-iyk} dy$$

where $k \in \mathbb{R}$ denotes the frequency, we can reduce the PDE problems in (2.2) and (2.3) to ODE problems. In each subdomain we have that

$$-\alpha_i \Delta u_i^n = 0 \qquad \text{in} \qquad \Omega_i. \tag{2.4}$$

Dropping the subscript i's and superscript n's, to simplify the notation, multiplying the PDE in (2.4) by e^{-iyk} and integrating from $-\infty$ to ∞ gives

$$-\alpha \int_{-\infty}^{\infty} e^{-iky} \Delta u \, dy = -\alpha \int_{\infty}^{\infty} e^{-iky} u_{xx} dy - \alpha \int_{\infty}^{\infty} e^{-iky} u_{yy} dy \tag{2.5}$$

Consider the second term on the right hand side of (2.5) involving u_{yy}. We have, using

integration by parts,

$$\begin{aligned}
-\alpha \int_{-\infty}^{\infty} e^{-iky} u_{yy} dy &= -\alpha \left[e^{-iky} u_y \right]_{y=-\infty}^{\infty} - \alpha \int_{-\infty}^{\infty} ike^{-iky} u_y dy \\
&= -\alpha \left[ike^{-iky} u \right]_{-\infty}^{\infty} + \alpha k^2 \int_{-\infty}^{\infty} e^{-iky} u dy \\
&= \alpha k^2 \hat{u}(x,k)
\end{aligned}$$

Next for the first term on the right hand side of (2.5) involving u_{xx} we have, using the Leibniz integral rule and the classical definition of a derivative, that

$$\begin{aligned}
-\alpha \int_{-\infty}^{\infty} e^{-iky} u_x dy &= -\alpha \frac{\partial}{\partial x} \left[\int_{-\infty}^{\infty} e^{-iky} u \, dy \right] \\
&= -\alpha \lim_{h \to 0} \frac{1}{h} \left[\int_{-\infty}^{\infty} u(x+h,y) e^{-iky} dy - \int_{-\infty}^{\infty} u(x,y) e^{-iky} dy \right] \\
&= -\alpha \lim_{h \to 0} \frac{1}{h} \left[\hat{u}(x+h,k) - \hat{u}(x,k) \right] \\
&= -\alpha \hat{u}_x(x,k).
\end{aligned}$$

A similar calculation yields that

$$-\alpha \int_{-\infty}^{\infty} e^{-iky} u_{xx} dy = -\alpha \hat{u}_{xx}(x,k).$$

Hence applying the Fourier transform to Laplace's equation in (2.4) gives

$$\alpha_i (k^2 - \partial_{xx}) \hat{u}_i^n(x,k) = 0.$$

The Fourier transformed OSM iteration is:

$$\begin{cases} \alpha_1(k^2 - \partial_{xx})\hat{u}_1^{n+1} = 0 \qquad \text{for} \quad x < 0, \quad k \in \mathbb{R} \\ (p_1 + \alpha_1\partial_x)\hat{u}_1^{n+1}(0,k) = (p_1 + \alpha_2\partial_x)\hat{u}_2^{n}(0,k) \qquad \text{for} \quad k \in \mathbb{R} \end{cases} \tag{2.6}$$

$$\begin{cases} \alpha_1(k^2 - \partial_{xx})\hat{u}_1^{n+1} = 0 \qquad \text{for} \quad x > 0, \quad k \in \mathbb{R} \\ (p_2 - \alpha_2\partial_x)\hat{u}_2^{n+1}(0,k) = (p_1 - \alpha_1\partial_x)\hat{u}_1^{n}(0,k) \qquad \text{for} \quad k \in \mathbb{R} \end{cases} \tag{2.7}$$

where the solution decays to zero at $\pm\infty$. The characteristic equations for the above ODEs have roots $\lambda_\pm = \pm|k|$.

Taking into account the behaviour of the solution at infinity we see subdomain solutions of (2.6) and (2.7) are of the form

$$\hat{u}_1^n(x,k) = A_n e^{|k|x} \tag{2.8}$$

and

$$\hat{u}_2^n(x,k) = B_n e^{-|k|x} \tag{2.9}$$

where

$$A_n = \hat{u}_1^n(0,k) \qquad \text{and} \qquad B_n = \hat{u}_2^n(0,k).$$

Now we have that

$$\partial_x\hat{u}_1^n = |k|A_n e^{|k|x} \qquad \text{and} \qquad \partial_x\hat{u}_2^n = -|k|B_n e^{-|k|x}$$

and the transmission conditions from (2.6) give us

$$(p_1 + \alpha_1|k|)A_n = (p_1 - \alpha_2|k|)B_{n-1}.$$

Hence

$$A_n = \frac{p_1 - \alpha_2|k|}{p_1 + \alpha_1|k|}B_{n-1}.$$

Similarly the transmission conditions from (2.7) give

$$B_n = \frac{p_2 - \alpha_1|k|}{p_2 + \alpha_2|k|}A_{n-1}.$$

In matrix vector form we have

$$\begin{pmatrix} A_n \\ B_n \end{pmatrix} = \begin{pmatrix} 0 & \frac{p_1-\alpha_2|k|}{p_1+\alpha_1|k|} \\ \frac{p_2-\alpha_1|k|}{p_2+\alpha_2|k|} & 0 \end{pmatrix} \begin{pmatrix} A_{n-1} \\ B_{n-1} \end{pmatrix}$$

Applying this again gives

$$\begin{pmatrix} A_n \\ B_n \end{pmatrix} = \begin{pmatrix} \frac{p_1-\alpha_2|k|}{p_1+\alpha_1|k|}\frac{p_2-\alpha_1|k|}{p_2+\alpha_2|k|} & 0 \\ 0 & \frac{p_1-\alpha_2|k|}{p_1+\alpha_1|k|}\frac{p_2-\alpha_1|k|}{p_2+\alpha_2|k|} \end{pmatrix} \begin{pmatrix} A_{n-2} \\ B_{n-2} \end{pmatrix}$$

Now plugging into (2.8) and (2.9) we have

$$\hat{u}_1^n(x,k) = \left(\frac{p_1 - \alpha_2|k|}{p_1 + \alpha_1|k|}\right)\left(\frac{p_2 - \alpha_1|k|}{p_2 + \alpha_2|k|}\right)\hat{u}_1^{n-2}(0,k)e^{|k|x}$$

and

$$\hat{u}_2^n(x,k) = \left(\frac{p_1 - \alpha_2|k|}{p_1 + \alpha_1|k|}\right)\left(\frac{p_2 - \alpha_1|k|}{p_2 + \alpha_2|k|}\right)\hat{u}_2^{n-2}(0,k)e^{-|k|x}$$

By induction

$$\hat{u}_1^{2n}(0,k) = \rho^n \hat{u}_1^0(0,k)$$

and

$$\hat{u}_2^{2n}(0,k) = \rho^n \hat{u}_2^0(0,k)$$

where the *convergence factor* of the OSM is given by

$$\rho = \rho(k, p_1, p_2) = \left| \frac{(p_1 - \alpha_2|k|)(p_2 - \alpha_1|k|)}{(p_1 + \alpha_1|k|)(p_2 + \alpha_2|k|)} \right| \tag{2.10}$$

In what follows we assume that the diffusion coefficient $\alpha_2 < \alpha_1$. The following result gives sufficient conditions for the Robin parameters under which the OSM iteration will converge.

Theorem 2.1.1. *Let $\alpha_2 < \alpha_1$. If $0 < p_1 \leq p_2$, for all $k \neq 0$ the optimised Schwarz iteration (2.2-2.3) converges, i.e. $\rho(k, p_1, p_2) < 1$.*

Proof. For the OSM iteration to converge we need

$$\left| \frac{(p_1 - \alpha_2|k|)(p_2 - \alpha_1|k|)}{(p_1 + \alpha_1|k|)(p_2 + \alpha_2|k|)} \right| < 1,$$

which is equivalent to the inequalities

$$-(p_1 + \alpha_1|k|)(p_2 + \alpha_2|k|) < (p_1 - \alpha_2|k|)(p_2 - \alpha_1|k|) < (p_1 + \alpha_1|k|)(p_2 + \alpha_2|k|).$$

Taking the inequality on the right first, expanding out the expressions we see that the inequality holds if and only if

$$p_1 + p_2 > 0. \tag{2.11}$$

For the leftmost inequality expanding out the expressions gives us

$$0 < 2p_1p_2 + |k|(p_1 - p_2)(\alpha_2 - \alpha_2) + 2\alpha_1\alpha_2k^2.$$

Then for the above inequality to hold it is enough if

$$p_1 p_2 \geq 0 \qquad \text{and} \qquad (p_1 - p_2)(\alpha_2 - \alpha_1) \geq 0.$$

Now since $\alpha_2 < \alpha_1$ combining these conditions with condition (2.11) it follows that we must have $0 < p_1 \leq p_2$. □

2.2 Optimised Robin parameters

We have seen that under suitable conditions the OSM will converge but we wish to choose the Robin parameters p_1 and p_2 to make the convergence as fast as possible. The obvious choices would be

$$p_1 = \alpha_2 |k| \qquad \text{and} \qquad p_2 = \alpha_1 |k|,$$

then the convergence factor would be identically zero and the iteration would converge in two steps. However $|k|$ is a frequency parameter that can take any value in $\mathbb{R}$, so when we back transform using the inverse Fourier transform parameters p_1 and p_2 will no longer be constants but complicated functions that would be difficult to implement. Instead we look for parameters p_1, $p_2 \in \mathbb{R}$ while uniformly optimising the convergence factor over a relevant range of frequencies. Then we consider the min-max problem:

$$\min_{p1,p2 \in \mathbb{R}} \left(\max_{k_{\min} \leq k \leq k_{\max}} \rho(k, p_1, p_2) \right) \tag{2.12}$$

The convergence factor (2.10) is symmetric about zero so we need only consider $k > 0$ rather than $|k|$ for (2.12). In the continuous case this still leaves a range of frequencies k from zero to infinity but since we want to implement the OSM in the discrete case we can give values for $k_{\min}$ and $k_{\max}$. Say we have performed a finite element discretisation

of Ω with mesh spacing h and where the interface Γ is of length H. The Nyquist-Shannon sampling criterion states that, to resolve a wave of frequency k one must have at least $2k$ samples on the interval $[0, 2\pi]$. Then as the lowest frequency we wish to resolve is H we have $k_{\min} = \pi/H$ and as the highest frequency we wish to resolve is h, $k_{\max} = \pi/h$.

We follow the proofs in [11, 22] for *one-sided*, *scaled one-sided* and *two-sided* Robin parameters. In all three cases we follow the same procedure. First we restrict the range of parameters p_1 and p_2, next we find local maxima in the frequency k and finally we see how these local maxima behave as we vary the Robin parameters.

2.2.1 One-sided Robin parameters

We first consider the simplest case when we have the same Robin parameter on either side of the artificial interface Γ. Setting

$$p_1 = p_2 = q,$$

the convergence factor becomes:

$$\rho(k, q) = \left| \frac{(q - \alpha_2 k)(q - \alpha_1 k)}{(q + \alpha_2 k)(q + \alpha_1 k)} \right| \tag{2.13}$$

and we wish to solve the min-max problem

$$\min_{q>0} \left(\max_{k_{\min} \leq k \leq k_{\max}} \rho(k, q) \right). \tag{2.14}$$

First we restrict the range of values that the Robin parameter q can take

Lemma 2.2.1. (Restricting the range of q) *For the min-max problem given by (2.14) we can restrict the range of q to $q \in [\alpha_2 k_{\min}, \alpha_1 k_{\max}]$.*

Proof. By assumption $q > 0$, then suppose $q \notin [\alpha_2 k_{\min},\, \alpha_1 k_{\max}]$, it follows that

$$(q - \alpha_2 k)(q - \alpha_1 k) > 0,$$

for all $k \in [k_{\min},\, k_{\max}]$. Then we can drop the absolute values from the convergence factor (2.13) and now consider

$$\rho(k, q) = \frac{(q - \alpha_2 k)(q - \alpha_1 k)}{(q + \alpha_2 k)(q + \alpha_1 k)}.$$

Taking the partial derivative of $\rho(k, q)$ with respect to q we have that

$$\frac{\partial \rho}{\partial q} = \frac{2k(\alpha_1 + \alpha_2)(q^2 - \alpha_1 \alpha_2 k^2)}{(q + \alpha_1 k)^2 (q + \alpha_2 k)^2}.$$

Now if $q < \alpha_2 k_{\min}$ we have that $q^2 < \alpha_2^2 k_{\min}^2 < \alpha_1 \alpha_2 k_{\min}^2$ and so $\frac{\partial \rho}{\partial q} < 0$ for all $k \in [k_{\min},\, k_{\max}]$. Hence increasing q will uniformly decrease $\rho(k, q)$ on $[k_{\min},\, k_{\max}]$. Then we must have that $q \geq \alpha_2 k_{\min}$.

If, on the other hand, $q > \alpha_1 k_{\max}$ we have that $q^2 > \alpha_1^2 k_{\max}^2 > \alpha_1 \alpha_2 k_{\max}^2$ and $\frac{\partial \rho}{\partial q} > 0$ for all $k \in [k_{\min},\, k_{\max}]$. Hence decreasing q will uniformly decrease $\rho(k, q)$. Hence we must have that $q \leq \alpha_1 k_{\max}$. □

Next we look for the local maxima of $\rho(k, q)$ as a function of k.

Lemma 2.2.2. (Local maxima in k) *Let $k_c = \frac{q}{\sqrt{\alpha_1 \alpha_2}}$. For a fixed q, the local maxima of $\rho(k, q)$ in the interval $[k_{\min},\, k_{\max}]$ are given by*

$$\max_{k_{\min} \leq k \leq k_{\max}} \rho(k, q) = \begin{cases} \max\{\rho(k_{\min}, q),\, \rho(k_c, q),\, \rho(k_{\max}, q)\} & \text{if} \quad k_c \in [k_{\min}, k_{\max}] \\ \max\{\rho(k_{\min}, q),\, \rho(k_{\max}, q)\} & \text{if} \quad k_c \notin [k_{\min}, k_{\max}]. \end{cases}$$

Proof. Taking the partial derivative of $\rho(k,q)$ with respect to k gives us

$$\frac{\partial \rho}{\partial k} = \frac{2q(\alpha_1+\alpha_2)(\alpha_1\alpha_2 k^2 - q^2)}{(q+\alpha_1 k)^2(q+\alpha_2 k)^2}.$$

Then $\frac{\partial \rho}{\partial k} = 0$ when

$$k = k_c = \frac{q}{\sqrt{\alpha_1\alpha_2}}.$$

By observing the sign of $\frac{\partial \rho}{\partial k}$ as we vary k near the point k_c, for a fixed q, we observe that

$$\rho(k,q) \quad \text{is} \quad \begin{cases} \text{strictly decreasing} & \text{for} \quad k \in \left[0,\, \frac{q}{\alpha_1}\right) \cup \left(\frac{q}{\sqrt{\alpha_1\alpha_2}},\, \frac{q}{\alpha_2}\right) \\ \text{strictly increasing} & \text{for} \quad k \in \left(\frac{q}{\alpha_1},\, \frac{q}{\sqrt{\alpha_1\alpha_2}}\right) \cup \left(\frac{q}{\alpha_2},\, \infty\right). \end{cases}$$

It follows that there is a local maxima of $\rho(k,q)$ at $k = k_c$. The full result of the Lemma follows from the fact that for some values of q the point k_c does not lie in the interval $[k_{\min},\, k_{\max}]$. □

We now have all the information we need to find the optimised Robin parameter that minimises the min-max problem (2.14). First observe that for the local maxima at $k = k_c$ the convergence factor simplifies to

$$R_c = \rho(k_c,\, q) = \frac{(\sqrt{\alpha_1} - \sqrt{\alpha_2})^2}{(\sqrt{\alpha_1} + \sqrt{\alpha_2})^2},$$

which is independent of parameter q. In fact depending on the situation there may be a unique, two distinct or an interval of minimising Robin parameters.

Theorem 2.2.3. (Optimised Robin parameter: one-sided) *Let*

$$\omega = \frac{\alpha_1}{\alpha_2}, \qquad k_r = \frac{k_{\max}}{k_{\min}}$$

and

$$f(\omega) = \Big((\omega+1)^2 + (\omega-1)\sqrt{\omega^2+6\omega+1}\Big)(4\omega)^{-1}.$$

Then

(i) If $k_r \geq f(\omega)$ a minimising parameter for the convergence factor (2.13) is given by $q^* = \sqrt{\alpha_1\alpha_2 k_{\min}k_{\max}}$. This minimiser q^* is unique when $\rho(k, q^*) \geq R_c$. Otherwise the minimum of the convergence factor can also be obtained by choosing any q in a closed interval around q^*.

(ii) If $k_r \leq f(\omega)$ then the convergence factor (2.13) has two distinct minimisers obtained by solving the equation

$$\rho(k_{\min}, q^*) = \rho(k_{\max}, q^*)$$

in the intervals $[\alpha_2 k_{\min}, \sqrt{\alpha_1\alpha_2 k_{\min}}]$ and $[\alpha_1\alpha_2 k_{\max}, \alpha_1 k_{\max}]$, respectively. In particular the two minimisers are the positive roots of the biquadratic polynomial given by

$$q^4 + \Big(\alpha_1\alpha_2(k_{\min}^2 + k_{\max}^2) - k_{\min}k_{\max}(\alpha_1+\alpha_2)^2\Big)q^2 + (\alpha_1\alpha_2 k_{\min}k_{\max})^2.$$

Proof. Even though the one-sided Robin parameters, $p_1 = p_2 = q$ is the simplest choice we could take, the results of the theorem and its proof are the most complex of the choices of parameters we study in this chapter. As such we omit the proof and instead refer the reader to the source material for the full details, [11]. □

Figure 2.1 shows the different convergence factors that arise as the jump in coefficients varies.

As previously mentioned when we discretise the continuous problem we get values for the maximum and minimum frequencies, namely $k_{\min} = \pi/H$ and $k_{\max} = \pi/h$ where H is the length of the interface and h the mesh spacing. We are interested in the asymptotic

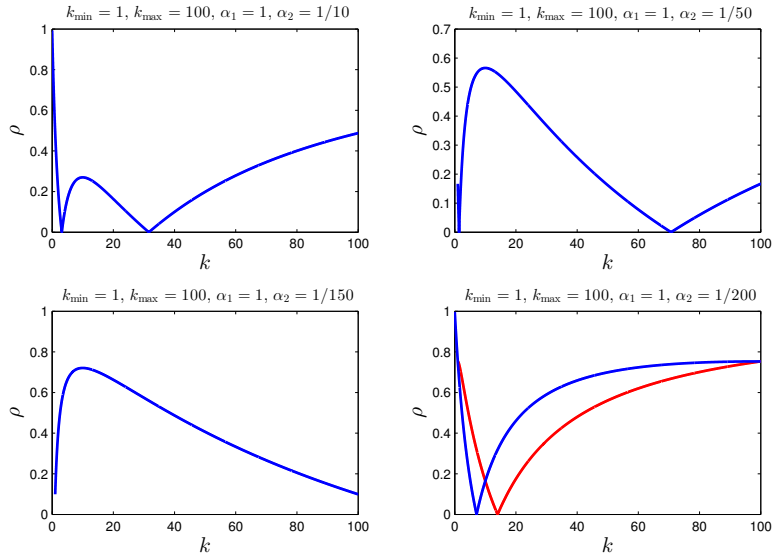

Figure 2.1: convergence factor with optimised one-sided Robin parameter

performance of the OSM when h tends to zero, i.e. when our approximation becomes more and more accurate to the continuous problem. Moreover we are interested in the asymptotic performance as the jump in diffusion coefficients becomes large, i.e. when α_2 tends to zero, where $\alpha_2 < \alpha_1$.

Theorem 2.2.4. (Asymptotic performance)

(i) When α_1 and α_2 are kept constant and h is small, with $k_{\max} = \pi/h$ the optimised one-sided Robin parameter is given by $q^* = \sqrt{\alpha_1\alpha_2 k_{\min}\pi}h^{-1/2}$. As $h \to 0$ the asymptotic convergence factor of the OSM is given by

$$\max_{k_{\min}\le k\le \pi/h} |\rho(k, q^*)| = 1 - 2\left(\frac{\alpha_1 + \alpha_2}{\sqrt{\alpha_1\alpha_2}}\right)\sqrt{\frac{k_{\min}}{\pi}}\sqrt{h} + O(h).$$

(ii) When α_1 is held constant and h is small and held constant, for small α_2 there are two distinct optimised one-sided Robin parameters, as set out in Theorem 2.2.3. For both optimised parameters, expanding the convergence factor $\rho(k, q^*)$ as $\alpha_2 \to 0$ and

keeping the dominant term when h is small, the asymptotic convergence factor of the OSM is given by

$$\max_{k_{\min}\leq k\leq \pi/h} |\rho(k,q^*)| = \frac{\sqrt{k_{\max}}-\sqrt{k_{\min}}}{\sqrt{k_{\max}}+\sqrt{k_{\min}}} + O\left(\frac{\alpha_2}{\alpha_1}\right) \approx 1 - 2\sqrt{\frac{k_{\min}h}{\pi}}.$$

Proof. (i) As $h \to 0$ and $k_{\max} = \pi/h$ becomes large, $k_r > f(\omega)$, giving us the minimiser $q^* = \sqrt{\alpha_1\alpha_2 k_{\min}}\pi h^{-1/2}$ for the convergence factor. Moreover as $h \to 0$, $\rho(k_{\min}, q^*) \to 1$. Now since R_c is a constant less than 1, for small enough h we have $\rho(k_{\min}, q^*) > R_c$ and the minimiser q^* is unique. Taylor expanding $\rho(k_{\min}, q^*)$ about $h = 0$ gives the result.

(ii) As $\alpha_2 \to 0$, $k_r < f(\omega)$ and we have two distinct minimisers given by the positive roots of the biquadratic in Theorem 2.2.3. Solving the biquadratic and expanding the roots we have

$$q_l^* = \sqrt{k_{\min}k_{\max}} + (k_{\max}-k_{\min})^2\left(\frac{\alpha_2}{\alpha_1}\right)^{-1} + O\left(\left(\frac{\alpha_2}{\alpha_1}\right)^{-2}\right)$$

and

$$q_r^* = \sqrt{k_{\min}k_{\max}}\left(\frac{\alpha_2}{\alpha_1}\right) - \frac{(k_{\max}-k_{\min})^2}{2\sqrt{k_{\min}k_{\max}}} + O\left(\left(\frac{\alpha_2}{\alpha_1}\right)^{-1}\right).$$

Plugging these into $\rho(k_{\min}, q^*)$ and expanding about $\alpha_2 = 0$ gives the same desired result for both choices of minimiser.

□

From the above we see for the choice of one-sided parameters we have an asymptotic performance of the convergence factor in h of $1 - O(\sqrt{h})$. This is consistent with the case of OSM in homogeneous media, [20]. There the optimised one-sided parameter is given by $q^* = \alpha\sqrt{k_{\min}k_{\max}}$, where $\alpha = \alpha_1 = \alpha_2$ and the asymptotics of the convergence factor is

again $1 - O(\sqrt{h})$.

In our heterogeneous case, as $\alpha_2 \to 0$, the convergence factor does not seem to depend on the jump asymptotically, rather the dominant term is a constant that depends on h, approaching 1 when h is small. We will see in our numerical experiments at the end of the chapter that the optimised one-sided parameters do not perform very well as the jump increases.

It is also possible to look at the asymptotics as both the jump is large, $\alpha_2 \to 0$, and the mesh size is small, $h \to 0$, simultaneously. This may be necessary in the presence of boundary layers, then we would choose the mesh size h to be a function of the jump α_2/α_1. This case is explored further in [22].

2.2.2 Scaled one-sided Robin parameters

Next we consider the case when we have the same Robin parameter q on either side of the artificial interface Γ but since we have a heterogeneous problem we scale the parameters to take into account the diffusion coefficient from the opposing subdomain. This has the added benefit of simplifying the convergence factor. Namely we let

$$p_1 = \alpha_2 q \qquad \text{and} \qquad p_2 = \alpha_1 q$$

and so the convergence factor reduces to

$$\rho(k,q) = \left| \frac{\alpha_1\alpha_2(q-k)^2}{(\alpha_2 q + \alpha_1 k)(\alpha_1 q + \alpha_2 k)} \right|.$$

Hence we wish to solve the min-max problem

$$\min_{q>0} \left(\max_{k_{\min} \le k \le k_{\max}} \rho(k,q) \right). \tag{2.15}$$

Note that since $\rho(k,q)$ is always positive we can ignore the absolute value sign above.

Lemma 2.2.5. (Restricting the range of q) *For the min-max problem (2.15) we can restrict the range of q to $q \in [k_{\min}, k_{\max}]$.*

Proof. Taking the partial derivative of $\rho(k,q)$ with respect to q we have

$$\frac{\partial \rho}{\partial q} = \frac{\alpha_1\alpha_2(\alpha_1+\alpha_2)^2 k(q+k)(q-k)}{(\alpha_2 q+\alpha_1 k)^2(\alpha_1 q+\alpha_2 k)^2}.$$

Now if $q < k_{\min}$, then $\frac{\partial \rho}{\partial q} < 0$, so increasing q will uniformly decrease ρ. Hence we must have that $q \geq k_{\min}$. On the other hand if $q > k_{\max}$, then $\frac{\partial \rho}{\partial q} > 0$ and decreasing q will uniformly decrease ρ. Hence we must have that $q \leq k_{\max}$. □

Lemma 2.2.6. (Local maxima in k) *The maxima of $\rho(k,q)$ on the interval $[k_{\min}, k_{\max}]$ can be computed by looking at only the end points, i.e.*

$$\max_{k_{\min}\leq k\leq k_{\max}} \rho(k,q) = \max\{\rho(k_{\min},q), \rho(k_{\max},q)\}.$$

Proof. Taking the partial derivative of $\rho(k,q)$ with respect to k we have

$$\frac{\partial \rho}{\partial k} = \frac{\alpha_1\alpha_2(\alpha_1+\alpha_2)^2 q(k+q)(k-q)}{(\alpha_2 q+\alpha_1 k)^2(\alpha_1 q+\alpha_2 k)^2}.$$

Then ρ has a stationary point at $k = q$, but observing the sign of $\frac{\partial \rho}{\partial k}$ as we vary k we see that

$$\rho(k,q) \quad \text{is} \quad \begin{cases} \text{strictly decreasing} & \text{for} \quad k \in [k_{\min}, q) \\ \text{strictly increasing} & \text{for} \quad k \in (q, k_{\max}] \end{cases}$$

which indicates a local minimum.

Hence $\rho(k,q)$ has no local maxima in the interval $(k_{\min},\ k_{\max})$ and instead has maxima at the end points $\rho(k_{\min},q)$ and $\rho(k_{\max},q)$..

□

Theorem 2.2.7. (Optimised Robin parameter: scaled one-sided) *The optimised convergence factor $\rho(k, q^*)$ must satisfy the equioscillation property*

$$\rho(k_{\min}, q^*) = \rho(k_{\max}, q^*),$$

which gives the unique optimised Robin parameter:

$$q^* = \sqrt{k_{\min} k_{\max}}.$$

Proof. Looking at the partial derivative of $\rho(k, q)$ with respect to q at the end points we see that

$$\frac{\partial \rho(k_{\min}, q)}{\partial q} > 0 \qquad \text{for all} \qquad q \in (k_{\min}, k_{\max}).$$

Hence $\rho(k_{\min}, q)$ is increasing with respect to q.

On the other hand

$$\frac{\partial \rho(k_{\max}, q)}{\partial q} < 0 \qquad \text{for all} \qquad q \in (k_{\min}, k_{\max}).$$

Hence $\rho(k_{\max}, q)$ is decreasing with respect to q.

Moreover

$$\rho(k_{\min}, k_{\min}) = \rho(k_{\max}, k_{\max}) = 0.$$

Hence $\rho(k, q)$ is minimised uniformly when its endpoints are equal, i.e. when

$$\rho(k_{\min}, q^*) = \rho(k_{\max}, q^*). \tag{2.16}$$

To see that $\rho(k, q)$ must satisfy the equioscillation property above consider if we had that

$\rho(k_{\min}, q) < \rho(k_{\max}, q)$. Since $\rho(k_{\max}, q)$ is decreasing with respect to q we can uniformly improve $\rho(k, q)$ by increasing q. On the other hand if we had that $\rho(k_{\min}, q) > \rho(k_{\max}, q)$, since $\rho(k_{\min}, q)$ is increasing with respect to q we can uniformly improve $\rho(k, q)$ by decreasing q. Hence we must have that $\rho(k_{\min}, q) = \rho(k_{\max}, q)$.

Solving (2.16) gives the unique Robin parameter:

$$q^* = \sqrt{k_{\min} k_{\max}}.$$

□

Theorem 2.2.8. (Asymptotic performance) *The scaled one-sided optimised Robin parameter is* $q^* = \sqrt{k_{\min} k_{\max}}$.

(i) When α_1 *and* α_2 *are kept constant,* h *is small and* $k_{\max} = \pi/h$ *the asymptotic convergence factor of the OSM with scaled one-sided Robin parameters as* $h \to 0$ *is given by*

$$\max_{k_{\min} \le k \le \pi/h} \rho(k, q^*) = 1 - \frac{(\alpha_1 + \alpha_2)^2}{\alpha_1 \alpha_2} \sqrt{\frac{k_{\min}}{\pi}} \sqrt{h} + O(h).$$

(ii) When α_1 *is held constant and* h *is small and held constant, for small* α_2*, expanding the convergence factor* $\rho(k, q^*)$ *as* $\alpha_2 \to 0$ *and keeping the dominant term when* h *is small, the asymptotic convergence factor of the OSM with scaled one-sided Robin parameters is given by*

$$\max_{k_{\min} \le k \le \pi/h} |\rho(k, q^*)| = \left(\frac{(\sqrt{k_{\max}} - \sqrt{k_{\min}})^2}{\sqrt{k_{\min} k_{\max}}} \right) \left(\frac{\alpha_2}{\alpha_1} \right) + O\left(\left(\frac{\alpha_2}{\alpha_1} \right)^2 \right) \approx \sqrt{\frac{\pi}{k_{\min} h}} \left(\frac{\alpha_2}{\alpha_1} \right). \tag{2.17}$$

Proof. Both results are obtained by plugging $q^* = \sqrt{k_{\min} k_{\max}}$ into $\rho(k, q^*)$ and Taylor expanding first about $h = 0$ and then $\alpha_2 = 0$. □

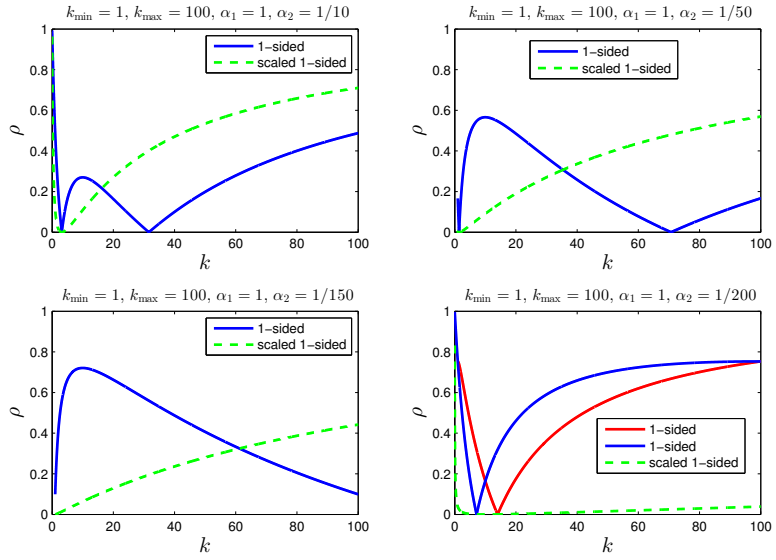

Figure 2.2: comparison of convergence factors for one-sided and scaled one-sided Robin parameters

For the optimised scaled one-sided Robin parameter we again have an asymptotic performance of the convergence factor when $h \to 0$ of $1 - O(\sqrt{h})$. The difference occurring between the one-sided and scaled one-sided choice of parameter is the asymptotic performance as the jump in coefficients becomes larger. We see from (2.17) that the convergence improves asymptotically as the jump becomes large. As $\alpha_2 \to 0$ the dominant term of the convergence factor is of the form $C(h)\frac{\alpha_2}{\alpha_1}$, where $C(h)$ is a constant that depends on the mesh size h. We will see in the numerical experiments at the end of the chapter that, for a fixed h, increasing the jump in coefficients leads to faster convergence of the OSM with this choice of parameter. This is an improvement from other popular DDMs such as FETI-DP and Neumann-Neumann that are independent of jumps in the coefficients, [50].

Figure 2.2 compares the convergence factor for the two choices of Robin parameter. We see that the choice of scaled one-sided parameter performs better as the jump in coefficients becomes larger.

2.2.3 Two-sided Robin parameters

Since we are considering a heterogeneous problem with different diffusion coefficients in the subdomains it seems wise to use two independent Robin parameters on each side of the interface. Accordingly we set

$$p_1 = \alpha_2 q_1 \qquad \text{and} \qquad p_2 = \alpha_1 q_2$$

and consider the convergence factor

$$\rho(k, q_1, q_2) = \frac{\alpha_1\alpha_2(q_1 - k)(q_2 - k)}{(\alpha_2 q_1 + \alpha_1 k)(\alpha_1 q_2 + \alpha_2 k)}$$

Recall from Theorem 2.1.1 that since $\alpha_1 > \alpha_2$ to keep $\rho(k, q_1, q_2) < 1$ we must have $q_1 \leq q_2$ and so we wish to solve the min-max problem

$$\min_{q_2 \geq q_1 > 0} \left(\max_{k_{\min} \leq k \leq k_{\max}} |\rho(k, q_1, q_2)| \right) \tag{2.18}$$

Note that since $\rho(k,\, q_1,\, q_2)$ may take negative values we must be mindful of the absolute value in (2.18).

Lemma 2.2.9. (Restricting the range of q_1 and q_2) *For the min-max problem (2.18) we can restrict the range of q_1 and q_2 to $q_1, q_2 \in [k_{\min}, k_{\max}]$.*

Proof. Taking the partial derivatives of $\rho(k, q_1, q_2)$ with respect to q_1 and q_2 gives

$$\frac{\partial \rho}{\partial q_1} = \frac{\alpha_1\alpha_2(\alpha_1 + \alpha_2)k(q_2 - k)}{(\alpha_2 q_1 + \alpha_1 k)^2(\alpha_1 q_2 + \alpha_2 k)} \tag{2.19}$$

and

$$\frac{\partial \rho}{\partial q_2} = \frac{\alpha_1\alpha_2(\alpha_1 + \alpha_2)k(q_1 - k)}{(\alpha_2 q_1 + \alpha_1 k)(\alpha_1 q_2 + \alpha_2 k)^2} \tag{2.20}$$

Looking at (2.19) if $q_2 < k_{\min}$, then $\frac{\partial \rho}{\partial q_1} < 0$ and so increasing q_2 will uniformly decrease $\rho(k, q_2, q_2)$. Hence we must have $q_2 \geq k_{\min}$.

On the other hand if $q_2 > k_{\max}$, then $\frac{\partial \rho}{\partial q_1} > 0$ and so decreasing q_2 will uniformly decrease $\rho(k, q_2, q_2)$. Hence we must have $q_2 \leq k_{\max}$.

Looking at (2.20) a similar calculation yields $q_1 \in [k_{\min}, k_{\max}]$.

□

Lemma 2.2.10. (Local maxima in k) *The maxima of $|\rho(k, q_1, q_2)|$ on the interval $[k_{\min}, k_{\max}]$ can be found by looking only at three points, where*

$$\max_{k_{\min} \leq k \leq k_{\max}} |\rho(k, q_1, q_2)| = \max\{\rho(k_{\min}, q_1, q_2), \rho(\sqrt{k_{\min} k_{\max}}, q_1, q_2), \rho(k_{\max}, q_1, q_2)\}.$$

Proof. Taking the partial derivative of $\rho(k, q_1, q_2)$ with respect to k gives

$$\frac{\partial \rho}{\partial k} = \frac{\alpha_1 \alpha_2 (\alpha_1 + \alpha_2)(\alpha_2 q_1 + \alpha_1 q_2)(k^2 - q_1 q_2)}{(\alpha_2 q_1 + \alpha_1 k)^2 (\alpha_2 q_2 + \alpha_2 k)^2}.$$

Now

$$\rho(k, q_1, q_2) \quad \text{is} \quad \begin{cases} \text{strictly decreasing} & \text{for} \quad k \in [k_{\min}, \sqrt{q_1 q_2}) \\ \text{strictly increasing} & \text{for} \quad k \in (\sqrt{q_1 q_2}, k_{\max}] \end{cases}$$

Moreover $\rho(k, q_1, q_2) < 0$ for $k \in (q_2, q_1)$. Hence $|\rho(k, q_1, q_2)|$ has a local maxima at $k = \sqrt{q_1 q_2}$. Then we have that

$$|\rho(k, q_1, q_2)| \quad \text{is} \quad \begin{cases} \text{strictly decreasing} & \text{for} \quad k \in [k_{\min}, q_1) \cup (\sqrt{q_1 q_2}, q_2) \\ \text{strictly increasing} & \text{for} \quad k \in (q_1, \sqrt{q_1 q_2}) \cup (q_2, k_{\max}] \end{cases}$$

and so $|\rho(k, q_1, q_2)|$ has three local maxima at $\rho(k_{\min}, q_1, q_2)$, $\rho(\sqrt{k_{\min} k_{\max}}, q_1, q_2)$ and $\rho(k_{\max}, q_1, q_2)$.

□

Theorem 2.2.11. (Optimised Robin parameters: two-sided) *The optimised convergence factor $\rho(k, q_1^*, q_2^*)$ must satisfy the equioscillation property*

$$\rho(k_{\min}, q_1^*, q_2^*) = \rho(\sqrt{q_1^* q_2^*}, q_1^*.q_2^*) = \rho(k_{\max}, q_1^*, q_2^*).$$

Then the unique optimised two-sided Robin parameters q_1^* and q_2^* are found by solving the non-linear system:

$$q_1^* q_2^* = k_{\min} k_{\max}$$
$$\rho(k_{\min}, q_1^*, q_2^*) = \rho(\sqrt{q_1^* q_2^*}, q_1^*, q_2^*),$$

where $q_1^* \leq q_2^*$.

Proof. Consider first only the end points $\rho(k_{\min}, q_1, q_2)$ and $\rho(k_{\max}, q_1, q_2)$. Observing the partial derivatives of these endpoints with respect to q_1 and q_2 we see that $\rho(k_{\min}, q_1, q_2)$ is increasing in both q_1 and q_2 while $\rho(k_{\max}, q_1, q_2)$ is decreasing in both q_1 and q_2. Hence $\rho(k, q_1, q_2)$ is minimised uniformly when the endpoints are equal, i.e. when

$$\rho(k_{\min}, q_1, q_2) = \rho(k_{\max}, q_1, q_2).$$

Solving for q_2 the above gives

$$q_2 = \frac{k_{\min} k_{\max}}{q_1}. \tag{2.21}$$

Hence $q_1 q_2 = k_{\min} k_{\max}$ and we have that $k_{\min} \leq q_1^* \leq \sqrt{k_{\min} k_{\max}} \leq q_2^* \leq k_{\max}$.

Now plugging (2.21) into $\rho(k, q_1, q_2)$ we can reduce the min-max problem in two param-

eters to the one parameter min-max problem, where we have that $k_{\min} \leq q_1^* \leq \sqrt{k_{\min}k_{\max}}$:

$$\min_{k_{\min} \leq q_1^* \leq \sqrt{k_{\min}k_{\max}}} (\max\{|R_1(q_1)|, |R_2(q_1)|\})$$

here

$$R_1(q_1) = \rho(k_{\min}, \frac{k_{\min}k_{\max}}{q_1}, q_1)$$

and

$$R_2(q_1) = \rho(\sqrt{k_{\min}k_{\max}}, \frac{k_{\min}k_{\max}}{q_1}, q_1).$$

Looking at the partial derivatives of $R_1(q_1)$ and $R_2(q_1)$ with respect to q_1 we see that

$$\frac{dR_1}{dq_1} > 0 \qquad \text{for} \qquad q_1 \in (k_{\min}, \sqrt{k_{\min}k_{\max}}]$$

and

$$\frac{dR_2}{dq_1} < 0 \qquad \text{for} \qquad q_1 \in (k_{\min}, \sqrt{k_{\min}k_{\max}}].$$

Moreover for $q_1 = k_{\min}$

$$0 = R_1(k_{\min}) < R_2(k_{\min})$$

and for $q_1 = \sqrt{k_{\min}k_{\max}}$

$$R_1(\sqrt{k_{\min}k_{\max}}) > R_2(\sqrt{k_{\min}k_{\max}}) = 0.$$

Hence the convergence factor $\rho(k, q_1, q_2)$ is uniformly minimised when

$$R_1(q_1) = R_2(q_1).$$

□

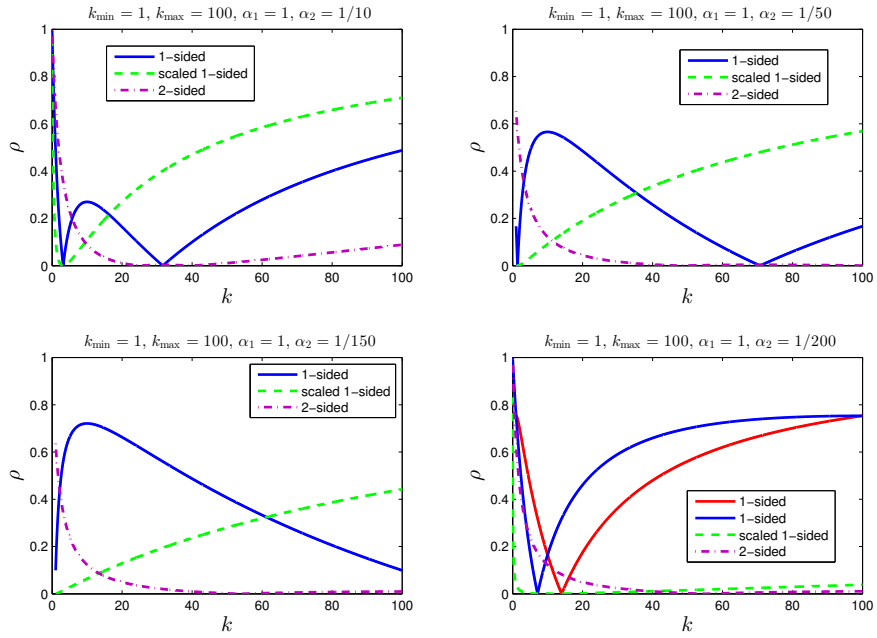

Figure 2.3: comparison of convergence factors for one-sided, scaled one-sided and two-sided Robin parameters

Instead of computing q_1^* and q_2^* by solving the non-linear system in the above theorem the task of computing q_1^* can be reduced to finding the roots of a biquadratic polynomial. Let $\omega = \alpha_2/\alpha_1$ then q_1^* is the unique real root in the interval $(k_{\min}, \sqrt{k_{\min}k_{\max}})$ of the following biquadratic:

$$(q_1 + \omega k_{\min})(q_1 + \omega k_{\max})(\sqrt{k_{\min}k_{\max}} - q_1)^2 - (q_1 - k_{\min})(k_{\max} - q_1)(q_1 + \omega\sqrt{k_{\min}k_{\max}})^2 = 0 \tag{2.22}$$

and q_2^* can be computed using the formula $q_2^* = (k_{\min}k_{\max})/q_1^*$.

Theorem 2.2.12. (Asymptotic performance)

(i) When α_1 and α_2 are kept constant, h is small and $k_{\max} = \pi/h$ the optimised two-sided

Robin parameters are

$$q_1^* \approx \frac{2\alpha_1 k_{\min}}{\alpha_1 - \alpha_2} - \frac{4k_{\min}^{3/2}}{\sqrt{\pi}} \frac{\alpha_1(\alpha_1+\alpha_2)^2}{(\alpha_1-\alpha_2)^3}\sqrt{h}$$

and

$$q_2^* \approx \frac{\pi(\alpha_1-\alpha_2)}{2\alpha_1} h^{-1} + \sqrt{k_{\min}\pi}\frac{(\alpha_1+\alpha_2)^2}{\alpha_1(\alpha_1-\alpha_2)} h^{-1/2}.$$

The asymptotic convergence factor of the OSM with two-sided Robin parameters as $h \to 0$ *is given by*

$$\max_{k_{\min}\le k\le \pi/h} |\rho(k, q_1^*, q_2^*)| = \frac{\alpha_2}{\alpha_1} - 4\frac{\alpha_2(\alpha_1+\alpha_2)}{\alpha_1(\alpha_1-\alpha_2)}\sqrt{\frac{k_{\min}}{\pi}}\sqrt{h} + O(h). \tag{2.23}$$

(ii) Let $q_1^0 = \lim_{\alpha_2\to 0} q_1^*$ *and* $q_2^0 = \lim_{\alpha_2\to 0} q_2^*$. *When* α_1 *is held constant and* h *is small and held constant, for small* α_2, *expanding the convergence factor* $\rho(k, q_1^*, q_2^*)$ *as* $\alpha_2 \to 0$ *and keeping the dominant term when* h *is small, the asymptotic convergence factor of the OSM with two-sided Robin parameters is given by*

$$\max_{k_{\min}\le k\le \pi/h} |\rho(k, q_1^*, q_2^*)| = \frac{(q_1^0 - k_{\min})(q_2^0 - k_{\min})}{k_{\min} q_2^0}\left(\frac{\alpha_2}{\alpha_1}\right) + O\left(\left(\frac{\alpha_2}{\alpha_1}\right)^2\right) \approx \frac{\alpha_2}{\alpha_1}.$$

Proof. (i) We establish the ansatz $q_1^* = C_1 h^{\beta_1} + C_2 h\beta_2$ for some $\beta_1 < \beta_2 \in \mathbb{R}$. Plugging the ansatz into the formula $R_1(q_1^*) = R_2(q_1^*)$ and expanding for small h we find that $\beta_1 = 0$ and $\beta_2 = 1/2$. The coefficients of the ansatz are

$$C_1 = \frac{2\alpha_1 k_{\min}}{\alpha_1 - \alpha_2}, \quad \text{and} \quad C_2 = \frac{4k_{\min}^{3/2}}{\sqrt{\pi}}\frac{\alpha_1(\alpha_1+\alpha_2)^2}{(\alpha_1-\alpha_2)^3}.$$

The formula for q_2^* is given by $q_2^* = (k_{\min}k_{\max})/q_1^*$. Now the result follows by expanding

$R_1(q_1^*)$ for small h.

(ii) Since both optimised two-sided parameters are contained within the interval $[k_{\min}, k_{\max}]$ they must be constant with respect to α_2 to leading order. Then using the leading order term from polynomial (2.22) we have

$$q_1^0 = \lim_{\alpha_2 \to 0} q_1^* = \frac{1}{4}\left(\sqrt{k_{\min}} + \sqrt{k_{\max}}\right) - \frac{1}{4}\sqrt{(\sqrt{k_{\min}} + \sqrt{k_{\max}})^4 - 16 k_{\min} k_{\max}}$$

$$q_2^0 = \lim_{\alpha_2 \to 0} q_2^* = \frac{1}{4}\left(\sqrt{k_{\min}} + \sqrt{k_{\max}}\right) + \frac{1}{4}\sqrt{(\sqrt{k_{\min}} + \sqrt{k_{\max}})^4 - 16 k_{\min} k_{\max}}$$

The result follows from expanding $\rho(k_{\min}, q_1^0, q_2^0)$ for small α_2.

□

For optimised two-sided parameters we have the surprising result that as $h \to 0$ the convergence factor doesn't deteriorate with the mesh parameter. The leading term α_2/α_1 in (2.23) is bounded away from 1 and as we increase the jump in coefficients convergence will improve. Asymptotically as h is fixed and $\alpha_2 \to 0$ we again observe improving convergence as we did for scaled one-sided parameters, though the two-sided parameters will perform better due to the smaller leading term in the expansion.

Figure 2.3 compares the optimised convergence factors for the three choices of Robin parameter. We observe that as the jump in coefficients is increased the scaled one-sided and two-sided parameters yield much smaller convergence factors than that of the one-sided parameter. Moreover the two-sided parameters perform better than the scaled one-sided parameters. This confirms our theoretical results which we will test with numerical examples in the last section of this chapter.

2.3 The discrete optimised Schwarz method

So far the optimised Schwarz methods we have considered are continuous. In practice we wish to approximate the solutions to problems using an appropriate discretisation method.

We assume the partition of the domain Ω into non-overlapping subdomains, Ω_1 and Ω_2, is such that for both subdomains $\partial\Omega_i \cap \partial\Omega \neq \emptyset$, i.e. we have no subdomains that "float". We start by deriving the weak formulation of (2.1).

Let $f \in L^2(\Omega)$ and consider test functions $v \in H^1(\Omega_i) \cap H_0^1(\Omega)$. Using the identity $\int_{\partial\Omega_i} = \int_{\partial\Omega_i \cap \partial\Omega} + \int_\Gamma$ and Green's identity we have from (2.1) that

$$\begin{aligned}
\int_{\Omega_i} f v \, dx &= -\int_{\Omega_i} a_i \Delta u_i^k v \, dx \\
&= \int_{\Omega_i} a_i \nabla u_i^k \cdot \nabla v \, dx - \int_{\partial\Omega_i} a_i \partial_{n_i} u_i^k v \, dS \\
&= \int_{\Omega_i} a_i \nabla u_i^k \cdot \nabla v \, dx - \int_{\partial\Omega_i \cap \partial\Omega} a_i \partial_{n_i} u_i^k v \, dS - \int_\Gamma a_i \partial_{n_i} u_i^k v \, dS \\
&= \int_{\Omega_i} a_i \nabla u_i^k \cdot \nabla v \, dx - \int_\Gamma (\lambda_i^{k-1} - p_i u_i^k) v \, dS.
\end{aligned}$$

Hence the weak formulation of (2.1) is to find $u_i^k \in H^1(\Omega_i) \cap H_0^1(\Omega)$ such that

$$\int_{\Omega_i} a_i \nabla u_i^k \cdot \nabla v \, dx + \int_\Gamma p_i u_i^k v \, dS = \int_{\Omega_i} f v \, dx + \int_\Gamma \lambda_i^{k-1} v \, dS, \tag{2.24}$$

for all $v \in H^1(\Omega_i) \cap H_0^1(\Omega)$. The weak form given by (2.24) is well defined through the requirement that both Ω_1 and Ω_2 are Lipschitz since, then Γ is Lipschitz and the surface integrals are defined. The Lipschitz requirement is necessary when we consider the discrete 2LM method and estimate the eigenvalues of Schur complement matrices S_1 and S_2 on page 90, which assumes Lipschitz polygonal domains. Extensions to other geometries are possible, such as fractal domains, see [66], though we do not consider them here. Next we make explicit the term involving λ_i^{k-1} on the right hand side of (2.24).

At the kth step of the OSM iteration we have that $-a_{3-i}\Delta u_{3-i}^{k-1} = f$ outside Ω_i. Let $\Omega_i^c = \Omega \setminus \Omega_i$ and let ∂n_i^c denote the directional derivative with respect to the outward pointing normal of Ω_i^c and so $\partial n_i^c = -\partial n_i$. A similar calculation as above on $-a_{3-i}\Delta u_{3-i}^{k-1} = f$ yields

$$\begin{aligned}\int_{\Omega_i^c} fv\,dx &= \int_{\Omega_i^c} a_{3-i}\nabla u_{3-i}^{k-1}\cdot\nabla v\,dx - \int_{\partial\Omega_i^c\cap\partial\Omega} a_{3-i}\partial_{n_i^c}u_{3-i}^{k-1}v\,dS - \int_\Gamma a_{3-i}\partial_{n_i^c}u_{3-i}^{k-1}v\,dS\\ &= \int_{\Omega_i^c} a_{3-i}\nabla u_{3-i}^{k-1}\cdot\nabla v\,dx + \int_\Gamma a_{3-i}\partial_{n_i}u_{3-i}^{k-1}v\,dS\\ &= \int_{\Omega_i^c} a_{3-i}\nabla u_{3-i}^{k-1}\cdot\nabla v\,dx + \int_\Gamma(\lambda_i^{k-1} - p_i u_{3-i}^{k-1})v\,dS.\end{aligned}$$

Hence

$$\int_\Gamma \lambda_i^{k-1}v\,dS = \int_{\Omega_i^c} fv\,dx - \int_{\Omega_i^c} a_{3-i}\nabla u_{3-i}^{k-1}\cdot\nabla v\,dx + \int_\Gamma p_i u_{3-i}^{k-1}v\,dS.$$

Substituting the above into (2.24) and using the identities $\int_\Omega = \int_{\Omega_i} + \int_{\Omega_i^c}$ and $\int_{\Omega_i^c} = \int_\Omega - \int_{\Omega_i}$ the weak formulation of (2.1) is now to find $u_i^k \in H^1(\Omega_i)\cap H_0^1(\Omega)$ such that

$$\begin{aligned}\int_{\Omega_i} a_i\nabla u_i^k\cdot\nabla v\,dx + \int_\Gamma p_i u_i^k v\,dS = \int_\Omega fv\,dx - \int_\Omega a_{3-i}\nabla u_{3-i}^{k-1}\cdot\nabla v\,dx\\ + \int_{\Omega_i} a_{3-i}\nabla u_{3-i}^{k-1}\cdot\nabla v\,dx + \int_\Gamma p_i u_{3-i}^{k-1}v\,dS,\end{aligned} \tag{2.25}$$

for all $v \in H^1(\Omega_i)\cap H_0^1(\Omega)$.

To derive the discrete form of (2.25) we perform a quasi-uniform triangulation of Ω, with mesh parameter h, into n degrees of freedom. The triangulation, $\mathcal{T}_h$, is such that each element lies in only one of the subdomains, so that the interface Γ does not "cut through" any elements. Once a suitable basis has been chosen for the finite element space $V_h \subset H_0^1(\Omega)$, we can construct the discretised form of (1.2) to obtain the linear system (1.1), with A an $n\times n$ sparse, symmetric positive definite matrix.

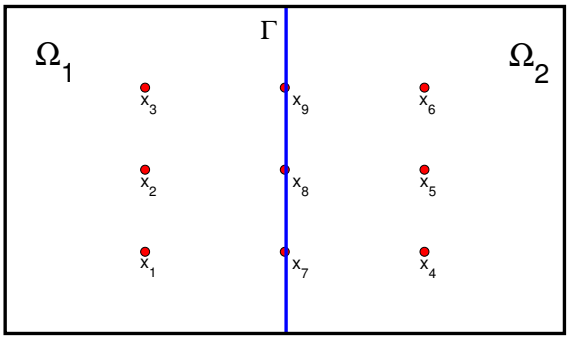

Figure 2.4: example of a discretised domain

For each subdomain Ω_i we can define a *restriction matrix* R_i. If the triangulation $\mathcal{T}_h$ contains n_i degrees of freedom in subdomain Ω_i and n_Γ degrees of freedom on the interface Γ we let $m_i = n_i + n_\Gamma$. Then R_i is an $m_i \times n$ Boolean matrix that restricts an arbitrary n dimensional vector $\boldsymbol{u}$ to an m_i dimensional vector $R_i\boldsymbol{u}$ which contains only the entries of $\boldsymbol{u}$ corresponding to degrees of freedom in $\Omega_i \cup \Gamma$. The corresponding extension matrix R_i^T prolongs an arbitrary m_i dimensional vector to an n dimensional vector.

For example in Figure 2.4 with the nodes numbered such that those in Ω_1 are first then those for Ω_2 and then those on Γ, the restriction matrix R_1 would be

$$R_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Using these restriction and extension matrices we can recover the global stiffness matrix

and load vector for (1.1) through the relations

$$A = \sum_{i=1}^{2} \alpha_i R_i^T A_{N_i} R_i \qquad \text{and} \qquad \boldsymbol{f} = \sum_{i=1}^{2} R_i^T \boldsymbol{f}_i. \tag{2.26}$$

Here A_{N_i} are local stiffness matrices for the subdomains with entries given by

$$(A_{N_i})_{jk} = \int_{\Omega_i} \alpha_0(\boldsymbol{x}) \left(\nabla\phi_j \cdot \nabla\phi_k\right) \, dx,$$

where $\phi_1, \ldots, \phi_n$ are the basis functions of the finite element space V_h. These matrices correspond to the discretisation of the Laplacian on subdomain Ω_i with Dirichlet boundary conditions on $\partial\Omega_i \setminus \Gamma$ and Neumann boundary conditions on Γ. The entries of $\boldsymbol{f}_i$ are given by $\int_{\Omega_i} f\phi_k \, dx$.

To take into account the Robin transmission condition terms in (2.25) we next define the mass matrix M on the interface Γ with entries given by

$$(M)_{jk} = \int_{\Gamma} \phi_j \phi_k \, dS.$$

To simplify the implementation of the OSM it is judicious to lump this mass matrix. Then we can replace M with the spectrally equivalent matrix hI, where h is the finite element mesh parameter and I the $n_\Gamma \times n_\Gamma$ identity matrix corresponding to the n_Γ degrees of freedom on Γ, [56].

Now let $\boldsymbol{u}_i$ denote the restriction of the solution vector $\boldsymbol{u}$ to $\Omega_i \cup \Gamma$. Then, from (2.25), the discrete version of the optimised Schwarz method is, given initial guess $\boldsymbol{u}_i^0$, for

$k = 1, 2, \ldots$: solve for $i = 1, 2$

$$\Big(\alpha_i A_{N_i} + p_i B_i\Big)\boldsymbol{u}_i^k = R_i\Big(\boldsymbol{f} - \alpha_{3-i}\tilde{A}R_{3-i}^T\boldsymbol{u}_{3-i}^{k-1}\Big) + \Big(\alpha_{3-i}A_{N_i} + p_i B_i\Big)R_i R_{3-i}^T\boldsymbol{u}_{3-i}^{k-1}, \tag{2.27}$$

where B_i is a matrix with entries equal to zero corresponding to vertices interior to Ω_i and entries equal to the lumped mass matrix corresponding to vertices on Γ. The matrix $\tilde{A}$ is the global stiffness matrix without the contribution of the diffusion coefficients, i.e. $\tilde{A} = \sum_{i=1}^{2} R_i^T A_{N_i} R_i$.

2.4 Numerical experiments

We consider the model problem (1.2), with $f = 1$ and $\alpha_0(\boldsymbol{x}) = 1$, on the unit square partitioned into non-overlapping rectangular subdomains $\Omega_1 = (0, 0.5) \times (0, 1)$ and $\Omega_2 = (0.5, 1) \times (0, 1)$.

The current numerical experiments and those throughout the thesis are performed entirely in MATLAB without the use of outside packages. First an initial coarse uniform or quasi-uniform triangulation of the domain is described by a matrix that lists the points of the mesh and a matrix that describes the connections between them. A MATLAB script is run that takes this coarse triangulation and refines it a chosen number of times by splitting each triangle in the mesh into four. Once the triangulation has been refined to a desired level, with mesh parameter h, the point and connection matrices are passed to a MATLAB script that constructs the stiffness matrix and load vector for the discretised PDE.

For these numerical experiments a uniform triangulation of the unit square is performed with mesh parameter h and the PDE in (1.2) is discretised using piecewise linear, triangular

	$h = 1/16$	$h = 1/32$	$h = 1/64$	$h = 1/128$
$\omega = 10^{-1}$	12	13	20	30
$\omega = 10^{-2}$	35	39	38	37
$\omega = 10^{-3}$	41	60	85	119
$\omega = 10^{-4}$	46	66	95	135
$\omega = 10^{-5}$	50	72	103	139

Table 2.1: number of OSM iterations using optimised one-sided Robin parameters

finite elements. Let $\omega = \alpha_2/\alpha_1$, we fix α_1 and allow α_2 to vary. The stopping criterion for the OSM iteration is given by

$$\boldsymbol{e}^k = \|\boldsymbol{u} - \boldsymbol{u}^k\|_2,$$

where $\boldsymbol{u}$ is the solution to the FEM discretisation of the global problem and $\boldsymbol{u}^k$ the approximate solution to the global problem formed from combining the subdomain solutions $\boldsymbol{u}_1^k$ and $\boldsymbol{u}_2^k$ from step k of the OSM. As each $\boldsymbol{u}_1^k$ and $\boldsymbol{u}_2^k$ have entries corresponding to the interface, when they are combined to form $\boldsymbol{u}^k$ we only take the interface entries from $\boldsymbol{u}_1^k$. The OSM iteration is stopped when $\boldsymbol{e}^k < 10^{-8}$. The results for one-sided parameters are shown in Table 2.1, scaled one-sided parameters in Table 2.2 and two-sided parameters in Table 2.3.

Reading across the rows of the tables for the one-sided and scaled one-sided parameters, we see that for a fixed ω decreasing h and refining the mesh leads to an increase in the number of iterations. To confirm that the order of convergence is indeed $h^{-1/2}$ as observed in the theoretical results in Figure 2.5 we have a logarithmic plot of the number of iterations needed for convergence for a fixed jump $\omega = 10^{-3}$. The red line has a slope corresponding to a convergence rate of $h^{-1/2}$ and we observe that for both one-sided and scaled one-

	$h = 1/16$	$h = 1/32$	$h = 1/64$	$h = 1/128$
$\omega = 10^{-1}$	10	14	18	24
$\omega = 10^{-2}$	6	7	8	9
$\omega = 10^{-3}$	4	5	5	6
$\omega = 10^{-4}$	4	4	4	4
$\omega = 10^{-5}$	3	3	3	4

Table 2.2: number of OSM iterations using optimised scaled one-sided Robin parameters

	$h = 1/16$	$h = 1/32$	$h = 1/64$	$h = 1/128$
$\omega = 10^{-1}$	7	8	9	11
$\omega = 10^{-2}$	5	5	5	6
$\omega = 10^{-3}$	4	4	4	4
$\omega = 10^{-4}$	3	3	3	4
$\omega = 10^{-5}$	3	3	3	3

Table 2.3: number of OSM iterations using optimised two-sided Robin parameters

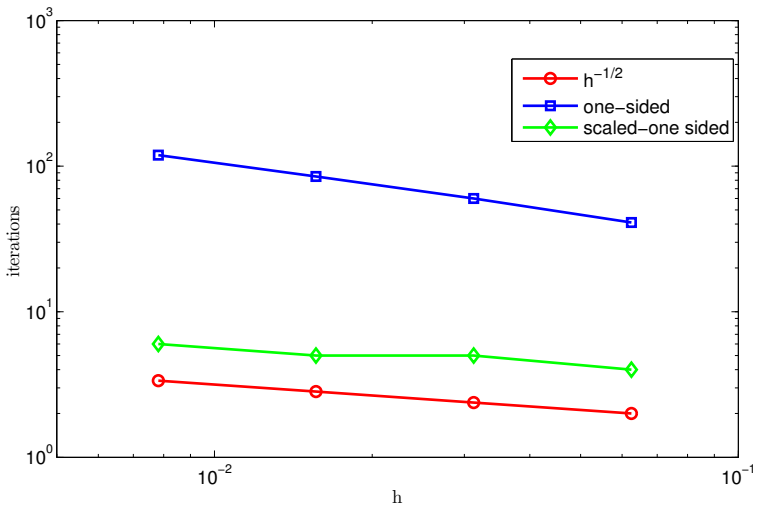

Figure 2.5: logarithmic plot of the number of iterations of the OSM with optimised one-sided and scaled one-sided parameters for different values of h, with $\omega = 10^{-3}$

sided parameters we achieve convergence that is at least this rate. In the case of two-sided parameters the theory stated that the convergence was independent of the mesh parameter h. In the numerical experiments we do observe a slight increase in the iteration count, but only when the jump in coefficients is small.

Reading down the columns of the tables gives us the results for a fixed h and a varying jump in the coefficients. We see that for the one-sided parameter the performance deteriorates as the jump increases. For scaled one-sided and two-sided parameters, on the other hand, the iteration count improves drastically, agreeing with the theory.

Although we did not perform the asymptotic analysis of taking the limit as the jump increases and the mesh becomes finer simultaneously we can observe some of this behaviour by starting at the top left and reading the tables diagonally down to the right. This corresponds to taking the limits $\omega \to 0$ and $h \to 0$ at various proportional rates $\frac{\omega}{h}$. Whereas for the one-sided parameters taking these limits simultaneously leads to worse iteration counts, for the scaled one-sided parameters there is a tension. Decreasing h leads to higher iteration counts while decreasing ω leads to fewer iterations. Depending on the

angle $\frac{\omega}{h}$ one or the other wins out.

Chapter 3

The two-Lagrange multiplier method

3.1 Formulation

The 2LM method is closely related to the OSM but does not directly introduce an iteration for system (1.1), rather a smaller equivalent system is solved. We consider the local Robin problems, for $i = 1, 2$

$$\begin{cases} -a_i \Delta u_i = f & \text{in} \quad \Omega_i \\ u_i = 0 & \text{on} \quad \partial\Omega_i \cap \partial\Omega \\ (p_i + a_i \partial_{n_i}) u_i = \lambda_i & \text{on} \quad \Gamma, \end{cases} \tag{3.1}$$

where the transmission conditions from subdomain Ω_{3-i} have been absorbed into the "Robin data" λ_i.

Following the formulation as presented in [63], after performing a quasi-uniform triangulation of the domain, we can write the local stiffness matrix A_{N_i} and load vector $\boldsymbol{f}_i$ in block form as

$$A_{N_i} = \begin{bmatrix} A_{II_i} & A_{I\Gamma_i} \\ A_{\Gamma I_i} & A_{\Gamma\Gamma_i} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{f}_i = \begin{bmatrix} \boldsymbol{f}_{I_i} \\ \boldsymbol{f}_{\Gamma_i} \end{bmatrix}.$$

Here the degrees of freedom have been partitioned into those internal to Ω_i and those on Γ. We can similarly partition the local solution vector into interior and interface blocks as $\boldsymbol{u}_i = [\boldsymbol{u}_{I_i}, \boldsymbol{u}_{\Gamma_i}]^T$. Then given a suitable "Robin data" vector $\boldsymbol{\lambda}_i$, the discrete form of (3.1) is given by

$$\begin{bmatrix} \alpha_i A_{II_i} & \alpha_i A_{I\Gamma_i} \\ \alpha_i A_{\Gamma I_i} & \alpha_i A_{\Gamma\Gamma_i} + p_i h I \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_{I_i} \\ \boldsymbol{u}_{\Gamma_i} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_{I_i} \\ \boldsymbol{f}_{\Gamma_i} \end{bmatrix} + \begin{bmatrix} 0 \\ \boldsymbol{\lambda}_i \end{bmatrix}, \tag{3.2}$$

where, as in the case of the OSM, the mass matrix has been lumped. We can eliminate the interior nodes of (3.2) to obtain

$$(\alpha_i S_i + p_i h I)\, \boldsymbol{u}_{\Gamma_i} = \boldsymbol{g}_i + \boldsymbol{\lambda}_i, \tag{3.3}$$

where

$$S_i = A_{\Gamma\Gamma_i} - A_{\Gamma I_i} A_{II_i}^{-1} A_{I\Gamma_i} \quad \text{and} \quad \boldsymbol{g}_i = \boldsymbol{f}_{\Gamma_i} - A_{\Gamma I_i} A_{II_i}^{-1} \boldsymbol{f}_{I_i},$$

are the *Schur complement* of the Neumann matrix A_{N_i} and the *condensed right hand side* of the load vector $\boldsymbol{f}_i$ respectively. The Schur complement is the discrete form of the Dirichlet to Neumann map and since A_{N_i} is symmetric positive definite the corresponding S_i matrix is symmetric positive definite. The assumption that neither subdomain "floats" ensures that the Schur complement matrices are non-singular.

Letting $p_s = \frac{(p_1+p_2)h}{2}$, (3.3) gives the relation

$$p_s \overbrace{\begin{bmatrix} \boldsymbol{u}_{\Gamma_1} \\ \boldsymbol{u}_{\Gamma_2} \end{bmatrix}}^{\boldsymbol{u}_G} = \overbrace{\begin{bmatrix} p_s\,(\alpha_1 S_1 + p_1 h I)^{-1} & 0 \\ 0 & p_s\,(\alpha_2 S_2 + p_2 h I)^{-1} \end{bmatrix}}^{Q} \left(\overbrace{\begin{bmatrix} \boldsymbol{g}_1 \\ \boldsymbol{g}_2 \end{bmatrix}}^{\boldsymbol{g}} + \overbrace{\begin{bmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \end{bmatrix}}^{\boldsymbol{\lambda}} \right). \tag{3.4}$$

Since we have two subdomains, the *many sided trace* $\boldsymbol{u}_G$ has pairs of entries for each degree of freedom on Γ and in general will correspond to a discontinuous function across Γ. For $\boldsymbol{u}_G$ to correspond to a continuous function the pairs of entries for each node on Γ must all

agree. To make this more precise consider the orthogonal projection matrix

$$K = \frac{1}{2}\begin{bmatrix} I & I \\ I & I \end{bmatrix},$$

where I is the $n_\Gamma \times n_\Gamma$ identity matrix, if n_Γ is the number of vertices on Γ. Matrix K acts to average function values for each degree of freedom on Γ and hence $\boldsymbol{u}_G$ corresponds to a continuous function across Γ when it satisfies the continuity condition:

$$K\boldsymbol{u}_G = \boldsymbol{u}_G \tag{3.5}$$

and the corresponding local solutions, $\boldsymbol{u}_i$, of (3.2) will meet continuously across Γ.

The 2LM method for (1.1) is to solve system (1.3) for $\boldsymbol{\lambda}$ where

$$A_{2LM} = (I - 2K)(Q - K) \quad \text{and} \quad \boldsymbol{c} = -(I - 2K)Q\boldsymbol{g}, \tag{3.6}$$

Vector $\boldsymbol{\lambda}$ is a many sided trace of Lagrange multipliers from which the method derives its name, as there are pairs of entries for each vertex on Γ. On solving (1.3) we have "Robin data" vectors $\boldsymbol{\lambda}_i$ that are plugged into the local problems (3.2), which can then be solved in parallel. The resulting local solutions $\boldsymbol{u}_i$ meet continuously across Γ and will "glue together" in a suitable way to give the global solution $\boldsymbol{u}$ of (1.1) such that $\boldsymbol{u}_i = R_i\boldsymbol{u}$.

Lemma 3.1.1. *Problem (1.3) is equivalent to (1.1).*

Proof. To recover the solution, $\boldsymbol{u}$, of (1.1) from that of (1.3), the "Robin data" vectors $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ must solve the local Robin problems (3.2), such that the local solutions $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ meet continuously across the interface Γ. So first we check that continuity condition (3.5)

holds. From (1.3) and (3.6) we have that

$$\begin{aligned}\boldsymbol{\lambda} &= -(Q-K)^{-1}(I-2K)^{-1}(I-2K)Q\boldsymbol{g} \\ &= -(Q-K)^{-1}Q\boldsymbol{g}. \end{aligned} \tag{3.7}$$

The above together with (3.4) gives us

$$\begin{aligned}\boldsymbol{u}_G &= \frac{1}{p_s}Q(\boldsymbol{g}-(Q-K)^{-1}Q\boldsymbol{g}) \\ &= \frac{1}{p_s}(I-Q(Q-K)^{-1})Q\boldsymbol{g} \\ &= \frac{1}{p_s}((Q-K)-Q)(Q-K)^{-1}Q\boldsymbol{g} \\ &= -\frac{1}{p_s}K(Q-K)^{-1}Q\boldsymbol{g}.\end{aligned}$$

Then since K is an orthogonal projection matrix:

$$\begin{aligned}K\boldsymbol{u}_G &= -\frac{1}{p_s}K^2(Q-K)^{-1}Q\boldsymbol{g} \\ &= -\frac{1}{p_s}K(Q-K)^{-1}Q\boldsymbol{g}.\end{aligned}$$

Hence $K\boldsymbol{u}_G = \boldsymbol{u}_G$ as required.

Imposing continuity across the interface is not sufficient, we must also ensure that the fluxes match. By the continuity condition there exists a unique $\boldsymbol{u}$ that restricts to $\boldsymbol{u}_i$:

$$\boldsymbol{u}_i = R_i\boldsymbol{u} \qquad \text{for } i = 1,\, 2. \tag{3.8}$$

If we impose on the solution $\boldsymbol{u}$ that equation (1.1) holds we obtain

$$\begin{aligned}
\boldsymbol{f} = A\boldsymbol{u} &= \sum_{i=1}^{2} \alpha_i R_i^T A_{N_i} R_i \boldsymbol{u} && \text{(using (2.26))} \\
&= \sum_{i=1}^{2} \alpha_i R_i^T A_{N_i} \boldsymbol{u}_i && \text{(using (3.8))} \\
&= \boldsymbol{f} + \sum_{i=1}^{2} R_i^T \begin{bmatrix} 0 \\ \boldsymbol{\lambda}_i - p_i h \boldsymbol{u}_{\Gamma_i} \end{bmatrix}. && \text{(using (2.26) and (3.2))}
\end{aligned}$$

Cancelling the $\boldsymbol{f}$ terms on each side, for the fluxes to match across the interface we need

$$\sum_{i=1}^{2} \boldsymbol{\lambda}_i - p_i h \boldsymbol{u}_{\Gamma_i} = 0. \tag{3.9}$$

As continuity condition (3.5) holds we have that $\boldsymbol{u}_{\Gamma_1} = \boldsymbol{u}_{\Gamma_2}$ and so $\boldsymbol{u}_G = [\boldsymbol{u}_{\Gamma_1}, \boldsymbol{u}_{\Gamma_1}]^T$. Then the left hand side of (3.9) becomes

$$\begin{aligned}
[I \quad I](\boldsymbol{\lambda} - p_s \boldsymbol{u}_G) &= [I \quad I](\boldsymbol{\lambda} - Q(\boldsymbol{g} + \boldsymbol{\lambda})) && \text{(using (3.4))} \\
&= [I \quad I]((K + (I - K) - Q)\boldsymbol{\lambda} - Q\boldsymbol{g}) \\
&= [I \quad I](-(Q - K)\boldsymbol{\lambda} - Q\boldsymbol{g} + (I - K)\boldsymbol{\lambda}) \\
&= [I \quad I](Q\boldsymbol{g} - Q\boldsymbol{g} + (I - K)\boldsymbol{\lambda}) && \text{(using (3.7))} \\
&= 0,
\end{aligned}$$

as required.

Now consider the opposite direction, we have a solution $\boldsymbol{u}$ to $A\boldsymbol{u} = \boldsymbol{f}$ and wish to recover the Robin data $\boldsymbol{\lambda}$ from the 2LM method. By construction this $\boldsymbol{u}$ meets continuously across the interface between the subdomains so we can recover the local solution $\boldsymbol{u}_i$ using (3.8). Next define the local Robin data $\boldsymbol{\lambda}_i$ using (3.2). Eliminating interior nodes and writing in

block form (3.2) becomes

$$p_s \boldsymbol{u}_G = Q(\boldsymbol{g} + \boldsymbol{\lambda}). \tag{3.10}$$

Since we have a solution $\boldsymbol{u}$ to $A\boldsymbol{u} = \boldsymbol{f}$, from (3.9) we also solve

$$\boldsymbol{\lambda} - p_s \boldsymbol{u}_G = 0.$$

Now left multiplying the above by the averaging matrix K we obtain

$$K\boldsymbol{\lambda} - p_s K \boldsymbol{u}_G = 0. \tag{3.11}$$

Subbing (3.10) into the continuity condition (3.5) and (3.11) we get

$$KQ(\boldsymbol{g} + \boldsymbol{\lambda}) = Q(\boldsymbol{g} + \boldsymbol{\lambda})$$

and

$$K\lambda - KQ(\boldsymbol{g} + \boldsymbol{\lambda}) = 0.$$

Adding the above two equations together we obtain

$$(Q - K)\boldsymbol{\lambda} = -Q\boldsymbol{g},$$

then left multiplying the above by $(I - 2K)$ we obtain the 2LM method:

$$(I - 2K)(Q - K)\boldsymbol{\lambda} = -(I - 2K)Q\boldsymbol{g}$$

as required. □

3.2 Equivalence to the optimised Schwarz method

We next show the equivalence between the OSM and the 2LM method in the absence of cross points when system (1.3) is solved using a *Richardson iteration.*

Lemma 3.2.1. *Let $\boldsymbol{\lambda}_i^k$ be generated by a Richardson iteration applied to (1.3):*

$$\boldsymbol{\lambda}^k = \boldsymbol{\lambda}^{k-1} + 2(\boldsymbol{c} - A_{2LM}\boldsymbol{\lambda}^{k-1}). \tag{3.12}$$

Let $\boldsymbol{u}_i^k$ be generated by the OSM iteration (2.27). Assume that $\boldsymbol{u}_i^k$ solves

$$\begin{bmatrix} \alpha_i A_{II_i} & \alpha_i A_{I\Gamma_i} \\ \alpha_i A_{\Gamma I_i} & \alpha_i A_{\Gamma\Gamma_i} + p_i hI \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_{I_i}^k \\ \boldsymbol{u}_{\Gamma_i}^k \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_{I_i} \\ \boldsymbol{f}_{\Gamma_i} + \boldsymbol{\lambda}_i^k \end{bmatrix}, \tag{3.13}$$

when $k = 0$. Then $\boldsymbol{u}_i^k$ solves (3.13) for all k.

Proof. First consider the iterates produced by (3.12):

$$\boldsymbol{\lambda}_i^k = 2p_s(\alpha_{3-i}S_{3-i} + p_{3-i}hI)^{-1}(\boldsymbol{g}_{3-i} + \boldsymbol{\lambda}_{3-i}^{k-1}) - \boldsymbol{\lambda}_{3-i}^{k-1}, \tag{3.14}$$

for $i = 1, 2$. Now consider the local Robin problem

$$\begin{bmatrix} \alpha_{3-i} A_{II_{3-i}} & \alpha_{3-i} A_{I\Gamma_{3-i}} \\ \alpha_{3-i} A_{\Gamma I_{3-i}} & \alpha_{3-i} A_{\Gamma\Gamma_{3-i}} + p_{3-i} hI \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_{I_{3-i}}^k \\ \tilde{\boldsymbol{u}}_{\Gamma_{3-i}}^k \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_{I_{3-i}} \\ \boldsymbol{f}_{\Gamma_{3-i}} + \boldsymbol{\lambda}_{3-i}^k \end{bmatrix}.$$

Eliminating the interior nodes and rearranging we have

$$\boldsymbol{\lambda}_{3-i}^k = (\alpha_{3-i}S_{3-i} + p_{3-i}hI)\tilde{\boldsymbol{u}}_{\Gamma_{3-i}}^k - \boldsymbol{g}_{3-i}. \tag{3.15}$$

Recall that $p_s = \frac{p_i+p_{3-i}}{2}h$. Then plugging (3.15) into (3.14) gives

$$\begin{aligned}\boldsymbol{\lambda}_i^k &= (p_i + p_{3-i})h\tilde{\boldsymbol{u}}_{3-i}^{k-1} - (\alpha_{3-i}S_{3-i} + p_{3-i}hI)\tilde{\boldsymbol{u}}_{\Gamma_{3-i}}^{k-1} + \boldsymbol{g}_{3-i} \\ &= -(\alpha_{3-i}S_{3-i} - p_ihI)\tilde{\boldsymbol{u}}_{\Gamma_{3-i}}^{k-1} + \boldsymbol{g}_{3-i}. \end{aligned} \tag{3.16}$$

Combining (3.15) and (3.16) we obtain the iteration:

$$(\alpha_iS_i + p_ihI)\tilde{\boldsymbol{u}}_{\Gamma_i}^k = -(\alpha_{3-i}S_{3-i} - p_ihI)\tilde{\boldsymbol{u}}_{\Gamma_{3-i}}^{k-1} + \boldsymbol{g}_{3-i} + \boldsymbol{g}_i. \tag{3.17}$$

Now consider the OSM iteration, (2.27), which in block form is given by

$$\begin{bmatrix} \alpha_iA_{II_i} & \alpha_iA_{I\Gamma_i} \\ \alpha_iA_{\Gamma I_i} & \alpha_iA_{\Gamma\Gamma_i} + p_ihI \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_{I_i}^k \\ \boldsymbol{u}_{\Gamma_i}^k \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_{I_i} \\ \boldsymbol{f}_\Gamma - \alpha_{3-i}A_{\Gamma I_{3-i}}\boldsymbol{u}_{I_{3-i}}^{k-1} - (\alpha_{3-i}A_{\Gamma\Gamma_{3-i}} - p_ihI)\boldsymbol{u}_{\Gamma_{3-i}}^{k-1} \end{bmatrix}, \tag{3.18}$$

where $\boldsymbol{f}_\Gamma = \boldsymbol{f}_{\Gamma_i} + \boldsymbol{f}_{\Gamma_{3-i}}$. On eliminating the interior nodes we have

$$(\alpha_iS_i + p_ihI)\boldsymbol{u}_{\Gamma_i}^k = \boldsymbol{g}_i + \boldsymbol{f}_{\Gamma_{3-i}} - \alpha_{3-i}A_{\Gamma I_{3-i}}\boldsymbol{u}_{I_{3-i}}^{k-1} - (\alpha_{3-i}A_{\Gamma\Gamma_{3-i}} - p_ihI)\boldsymbol{u}_{\Gamma_{3-i}}^{k-1}. \tag{3.19}$$

Now (3.18) gives that

$$\boldsymbol{u}_{I_{3-i}}^k = \frac{1}{\alpha_{3-i}}A_{II_{3-i}}^{-1}(\boldsymbol{f}_{I_{3-i}} - \alpha_{3-i}A_{I\Gamma_{3-i}}\boldsymbol{u}_{\Gamma_{3-i}}^k),$$

which substituting into (3.19) and rearranging yields the iteration:

$$(\alpha_iS_i + p_ihI)\boldsymbol{u}_{\Gamma_i}^k = -(\alpha_{3-i}S_{3-i} - p_ihI)\boldsymbol{u}_{\Gamma_{3-i}}^{k-1} + \boldsymbol{g}_{3-i} + \boldsymbol{g}_i. \tag{3.20}$$

Then provided iterations (3.17) and (3.20) have the same initial guess, i.e. $\tilde{\boldsymbol{u}}_{\Gamma_i}^0 = \boldsymbol{u}_{\Gamma_i}^0$,

they will produce the same iterates. □

Due to the equivalence established above, the Richardson iteration in (3.12) will converge if and only if the OSM converges, which is widely accepted for homogeneous problems, see [20, 23]. To prove that the Richardson iteration converges we would have to show that the spectral radius $\rho(I - 2A_{2LM}) < 1$. However as we can accelerate convergence to the solution of (1.3) by using a Krylov subspace method we focus our analysis on understanding these techniques instead of the Richardson iteration. Since the matrix A_{2LM} is non-symmetric we need to use a method such as GMRES whose speed of convergence we will examine.

Chapter 4

The GMRES method and its convergence

This chapter is a review of some well known results about GMRES and its convergence and does not contain any original work. However a good understanding of GMRES and its convergence is needed for the analysis we will perform on the 2LM method in later chapters. Our presentation of the formulation of GMRES and its practical implementation are quoted from the results given in [4], [54], [58] and [65].

4.1 Formulation

Given a linear system of the form:

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{4.1}$$

where $A \in \mathbb{R}^{n\times n}$ is an arbitrary non-singular, non-symmetric matrix, the *generalised minimal residual method* (GMRES), due to Saad and Schultz [59], is one of the iterative methods for solving system (4.1) known as *Krylov subspace methods*. A Krylov subspace

generated by linear system (4.1) is defined as:

$$\mathcal{K}_k = \text{span}\{\boldsymbol{r}^0,\, A\boldsymbol{r}^0,\, A^2\boldsymbol{r}^0,\, \ldots,\, A^{k-1}\boldsymbol{r}^0\},$$

where, given initial guess $\boldsymbol{x}^0$ to the solution $\boldsymbol{x}$ of (4.1), $\boldsymbol{r}^0 = \boldsymbol{b} - A\boldsymbol{x}^0$ is the *initial residual*. At step k of a Krylov subspace method iteration we approximate the exact solution $\boldsymbol{x}$ by finding a vector $\boldsymbol{x}^k \in \boldsymbol{x}^0 + \mathcal{K}_k$. What distinguishes different methods is the choice of error term that must be minimised while finding vector $\boldsymbol{x}^k$. In the case of GMRES the error term that must be minimised is the norm of the residual, given by $\|\boldsymbol{r}^k\|_2 = \|\boldsymbol{b} - A\boldsymbol{x}^k\|_2$. Here, for a given vector $\boldsymbol{x} \in \mathbb{R}^n$, $\|\boldsymbol{x}\|_2 = \sqrt{\sum_i^n |x_i|^2}$ is the vector 2-norm.

In order to represent the vectors $\boldsymbol{x}^k$ that we find in the GMRES method we need a suitable basis for Krylov subspace $\mathcal{K}_k$. The first obvious choice for a basis for $\mathcal{K}_k$ would be given by the vectors, $\boldsymbol{r}^0$, $A\boldsymbol{r}^0$, $\ldots$, $A^{k-1}\boldsymbol{r}^0$. However forming these vectors would be numerically costly. Computing large powers of the matrix A can lead to round off errors and the vectors may become close to linearly dependent. Instead of using this basis we want to compute a basis for $\mathcal{K}_k$ we can more easily utilise.

4.1.1 The Arnoldi Process

The *Arnoldi process* is a Gram-Schmidt like iteration that constructs an orthonormal basis for the Krylov subspace $\mathcal{K}_k$, with a matrix A that need not necessarily be symmetric. Algorithm 4.1 outlines the details of the process On completion of the process the vectors $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^{\boldsymbol{k}}$ form an orthonormal basis for the Krylov subspace $\mathcal{K}_k$.

In particular the Arnoldi process is a partial reduction of the matrix A to upper Hes-

Algorithm 4.1: The Arnoldi process

Let $\boldsymbol{v}^1 = \boldsymbol{r}^0/\|\boldsymbol{r}^0\|_2$.
for $j = 1, \ldots, k$ **do**
  $\boldsymbol{w}^j = A\boldsymbol{v}^j$.
  for $i = 1, \ldots, j$ **do**
    $h_{i,j} = (\boldsymbol{w}^j)^T\boldsymbol{v}^i$.
    $\boldsymbol{w}^j = \boldsymbol{w}^j - h_{i,j}\boldsymbol{v}^i$.
  end
  $h_{j+1,j} = \|\boldsymbol{w}^j\|_2$.
  $\boldsymbol{v}^{j+1} = \boldsymbol{w}^j/h_{j+1,j}$.
end

senberg form. A $(k+1) \times k$ *upper Hessenberg matrix* H_k is one of the form:

$$H_k = \begin{bmatrix} h_{11} & \cdots & & h_{1,k} \\ h_{21} & h_{22} & & \vdots \\ 0 & \ddots & \ddots & \\ \vdots & \ddots & h_{k,k-1} & h_{k,k} \\ 0 & \cdots & 0 & h_{k+1,k} \end{bmatrix}, \tag{4.2}$$

i.e. H_k is zero below the first subdiagonal. To reduce the $n \times n$ matrix A to upper Hessenberg form let V_k be the $n \times k$ matrix whose columns are the orthonormal basis vectors $\{\boldsymbol{v}^i\}$ computed in the Arnoldi process:

$$V_k = \left[\begin{array}{c|c|c|c} \boldsymbol{v}^1 & \boldsymbol{v}^2 & \cdots & \boldsymbol{v}^k \end{array} \right].$$

Moreover let H_k be the $(k+1) \times k$ upper Hessenberg matrix, as shown in (4.2), whose entries $h_{i,j}$ are again those computed in the Arnoldi process, then we have that

$$AV_k = V_{k+1}H_k. \tag{4.3}$$

Now since the columns of V_k and V_{k+1} are orthonormal we have the reduction of A to upper

Hessenberg form:

$$V_k^T A V_k = \tilde{H}_k,$$

where $\tilde{H}_k$ is the $k \times k$ upper Hessenberg matrix formed by deleting the last row of H_k.

The reduction is partial if in the Arnoldi algorithm $k < n$ and full or complete if $k = n$ where the matrix A is $n \times n$. We can now derive the main steps of the GMRES method.

4.1.2 The GMRES algorithm

Assume that k steps of the Arnoldi process have been performed and we have an orthonormal basis for the Krylov subspace $\mathcal{K}_k$. Then any vector $\boldsymbol{x}^k \in \boldsymbol{x}^0 + \mathcal{K}_k$ can be written in the form

$$\boldsymbol{x}^k = \boldsymbol{x}^0 + V_k \boldsymbol{y}^k, \tag{4.4}$$

where $\boldsymbol{y}^k \in \mathbb{R}^k$. The algorithm for the GMRES method finds a vector $\boldsymbol{y}^k$ such that the residual norm $\|\boldsymbol{r}^k\|_2 = \|\boldsymbol{b} - A\boldsymbol{x}^k\|_2$ is minimised.

Recall from the Arnoldi process that $\boldsymbol{r}^0 = \beta \boldsymbol{v}^1$ where $\beta = \|\boldsymbol{r}^0\|_2$, then in particular $\boldsymbol{r}^0 = \beta V_{k+1} \boldsymbol{e}^1$, where $\boldsymbol{e}^1 = [1, 0, \ldots, 0]^T$ is the first canonical basis vector for $\mathbb{R}^{k+1}$. Then from (4.3) and (4.4) we have that

$$\begin{aligned} \boldsymbol{b} - A\boldsymbol{x}^k &= \boldsymbol{b} - A(\boldsymbol{x}^0 + V_k \boldsymbol{y}^k) \\ &= \boldsymbol{r}^0 - V_{k+1} H_k \boldsymbol{y}^k \\ &= V_{k+1}(\beta \boldsymbol{e}^1 - H_k \boldsymbol{y}^k). \end{aligned}$$

Now since V_{k+1} is column wise orthogonal we can use the isometric property, that for any vector z, $\|V_{k+1} z\|_2 = \|z\|_2$, and we obtain

$$\|\boldsymbol{r}^k\|_2 = \|\beta \boldsymbol{e}^1 - H_k \boldsymbol{y}^k\|_2. \tag{4.5}$$

Then to minimise the residual norm above, the GMRES method algorithm involves solving the least squares problem for the overdetermined system $H_k\boldsymbol{y}^k = \beta\boldsymbol{e}^1$ using a QR decomposition. Once the minimising vector $\boldsymbol{y}^k$ has been found the norm of the residual, $\|\boldsymbol{r}^k\|_2$, can be compared against a given tolerance. We will see in the next subsection, on implementing the GMRES method, that by incrementally updating the QR decomposition we can solve the least squares problem and as a result automatically have access to the norm of the residual, while avoiding having to form and store the matrix H_k. The GMRES algorithm is repeated until the tolerance in the norm of the residual has been met and the approximate solution $\boldsymbol{x}^k$ for system (4.1) is calculated using (4.4). We note that the matrix V_{k-1} can be updated to V_k at each step of the method by amending to the end of V_k the column vector $\boldsymbol{v}^k$ that is calculated in the Arnoldi process. Formally the steps of the GMRES method are given in Algorithm 4.2.

Algorithm 4.2: The GMRES method

Given tolerance TOL in the residual norm and initial guess $\boldsymbol{x}^0$ compute $\boldsymbol{r}^0 = b - A\boldsymbol{x}^0$. Let $\boldsymbol{v}^1 = \boldsymbol{r}^0/\beta$, where $\beta = \|\boldsymbol{r}^0\|_2$.
for $k = 1, \ldots$ **do**
  $\boldsymbol{w}^k = A\boldsymbol{v}^k$.
  for $i = 1, \ldots, k$ **do**
    $h_{i,k} = (\boldsymbol{w}^k)^T\boldsymbol{v}^i$.
    $\boldsymbol{w}^k = \boldsymbol{w}^k - h_{i,k}\boldsymbol{v}^i$.
    $h_{k+1,k} = \|\boldsymbol{w}^k\|_2$.
    $\boldsymbol{v}^{k+1} = \boldsymbol{w}^k/h_{k+1,k}$.
  end
  Update the matrix V_k with $\boldsymbol{v}^{k+1}$.
  Incrementally update the QR decomposition for system $H_k\boldsymbol{y}^k = \beta\boldsymbol{e}^1$ to solve the least squares problem for $\boldsymbol{y}^k$.
  Exit loop if $\|\boldsymbol{r}^k\|_2 < TOL$.
end
Compute the approximate solution $\boldsymbol{x}^k = \boldsymbol{x}^0 + V_k\boldsymbol{y}^k$.

Apart from its speed, one of the main advantages of the GMRES method, compared to older iterative methods such as the stationary Richardson iteration, is that if the system

matrix is of size $n \times n$ we are guaranteed convergence in no more than n steps, assuming exact arithmetic. Furthermore the Gram-Schmidt orthonormalisation procedure of the Arnoldi process cannot break down unless the GMRES method has already converged to the exact solution.

4.1.3 Practical implementation of GMRES

In the implementation of the GMRES method, given in Algorithm 4.2, a large amount of the computation time is spent solving the least squares problem, $\min_{\boldsymbol{y}^k} \|\beta \boldsymbol{e}^1 - H_k \boldsymbol{y}^k\|_2$ in step 10. Another difficulty arising in the algorithm is that to test the stopping criterion we have to, at each step k, explicitly calculate the residual $\boldsymbol{r}^k$ and its norm. One approach, to reduce computation time, is to only calculate the residual at regular intervals to test for convergence. There is another approach that will automatically provide us with the norm of the residual as a result of an efficient way to solve the least squares problem. The idea is to compute a QR decomposition of system $H_k \boldsymbol{y}^k = \beta \boldsymbol{e}^1$, that can be update incrementally at each step k, using *Givens rotations*, [27].

For a given pair or indices j and k we define the Givens rotation matrix as

$$G_{jk} = \begin{bmatrix} 1 & 0 & \cdots & & & & & 0 & \\ 0 & \ddots & & & & & & \vdots & \\ \vdots & & 1 & & & & & & \\ & & & c_j & & s_j & & & \leftarrow \text{ row } j \\ & & & & \ddots & & & & \\ & & & -s_j & & c_j & & & \leftarrow \text{ row } k \\ & & & & & & 1 & & \\ \vdots & & & & & & & \ddots & 0 \\ 0 & \cdots & & & & & \cdots & 0 & 1 \end{bmatrix}$$

with $c_j^2 + s_j^2 = 1$. In particular the matrix G_{jk} is the identity matrix except for rows j

and k. The Givens rotation matrices are orthogonal and when left-multiplying a matrix or vector they allow, with a suitable choice of c_j and s_j, specified entries of the matrix or vector to be set equal to zero.

Recall that the $(k+1) \times k$ matrix H_k is upper Hessenberg. Then to form the QR decomposition of H_k we seek an orthogonal matrix Q_k such that the product $Q_k H_k$ is a matrix, R_k, in upper triangular form. To transform H_k to upper triangular form we must set to zero the entries of its first subdiagonal. We can achieve this by successively left-multiplying H_k by the $(k+1)\times(k+1)$ Givens rotation matrices $G_j = G_{j,j+1}$ for $j = 1, \ldots, k$. The scalars c_j and s_j are chosen in such a way that left-multiplying by the Givens matrix G_j will set to zero the entry $h_{j+1,j}$ on the first subdiagonal. Let $Q_k = G_k G_{k-1} \cdots G_1$ then we have that

$$Q_k H_k = R_k = \begin{bmatrix} \tilde{R}_k \\ 0 \end{bmatrix},$$

where $\tilde{R}_k$ is a $k \times k$ upper triangular matrix and so R_k is equal to $\tilde{R}_k$ except for an extra row of zeros. Similarly we define the vector

$$\boldsymbol{g}^k = \beta Q_k \boldsymbol{e}^1 = \begin{bmatrix} \tilde{\boldsymbol{g}}^k \\ \gamma_{k+1} \end{bmatrix},$$

where $\tilde{\boldsymbol{g}}^k = [\gamma_1, \ldots, \gamma_k]^T$. Now, using the fact that Q_k is orthogonal, to solve the least squares problem we must find the vector $\boldsymbol{y}^k$ that minimises

$$\begin{aligned} \|\beta \boldsymbol{e}^1 - H_k \boldsymbol{y}^k\|_2 &= \|Q_k \left(\beta \boldsymbol{e}^1 - H_k \boldsymbol{y}^k\right)\|_2 \\ &= \|\boldsymbol{g}^k - R_k \boldsymbol{y}^k\|_2 \\ &= \left\| \begin{bmatrix} \tilde{\boldsymbol{g}}^k \\ \gamma_{k+1} \end{bmatrix} - \begin{bmatrix} \tilde{R}_k \\ 0 \end{bmatrix} \boldsymbol{y}^k \right\|_2, \end{aligned}$$

which has solution $\boldsymbol{y}_*^k = \tilde{R}_k^{-1} \tilde{\boldsymbol{g}}^k$. We also observe from the above result that once we have

the minimal vector the norm of the residual is given by

$$\begin{aligned}\|\boldsymbol{r}^k\|_2 &= \|\beta \boldsymbol{e}^1 - H_k \boldsymbol{y}_*^k\|_2 \\ &= \left\| \begin{bmatrix} \tilde{\boldsymbol{g}}^k \\ \gamma_{k+1} \end{bmatrix} - \begin{bmatrix} \tilde{R}_k \\ 0 \end{bmatrix} \boldsymbol{y}_*^k \right\|_2 \\ &= |\gamma_{k+1}|,\end{aligned}$$

which is, in absolute value, the last element of the vector $\boldsymbol{g}^k$.

To see how we can update the QR decomposition incrementally assume that the GMRES method algorithm is already under way. At the start of step k we have the decomposition $Q_{k-1}^{(k)} H_{k-1} = R_{k-1}$ with

$$Q_{k-1}^{(k)} = G_{k-1}^{(k)} G_{k-2}^{(k)} \cdots G_1^{(k)},$$

here we have added the superscript to emphasise that $Q_{k-1}^{(k)}$ is a $k \times k$ matrix. Observe that the matrix H_k differs from H_{k-1} by only a row of mostly zeros and a column. After performing one more step of the Arnoldi process we have the column $\boldsymbol{h}^k = [(\tilde{\boldsymbol{h}}^k)^T, h_{k+1,k}]^T$, where $\tilde{\boldsymbol{h}}^k = [h_{1k}, \ldots, h_{kk}]^T$, that can be amended to H_{k-1} to obtain H_k:

$$H_k = \begin{bmatrix} H_{k-1} & \tilde{\boldsymbol{h}}^k \\ 0 & h_{k+1,k} \end{bmatrix},$$

Similarly we can amend the matrix $Q_{k-1}^{(k)}$, by a row and a column of zeros and set the last diagonal to equal one, to get

$$\tilde{Q}_{k-1}^{(k+1)} = \begin{bmatrix} Q_{k-1}^{(k)} & 0 \\ 0 & 1 \end{bmatrix}.$$

Let $\boldsymbol{\rho}^k = Q_{k-1}^{(k)} \tilde{\boldsymbol{h}}^k = [\rho_1, \ldots, \rho_k]^T$. Now we have that

$$\tilde{Q}_{k-1}^{(k+1)} H_k = \begin{bmatrix} R_{k-1} & \boldsymbol{\rho}^k \\ 0 & h_{k+1,k} \end{bmatrix}.$$

Then to obtain R_k from the above we must set to zero the last element, $h_{k+1,k}$, which we can achieve by left-multiplying using the Givens rotation matrix $G_k^{(k+1)}$ with entries

$$c_k = \frac{\rho_k}{\sqrt{\rho_k^2 + h_{k+1,k}^2}} \qquad \text{and} \qquad s_k = \frac{h_{k+1,k}}{\sqrt{\rho_k^2 + h_{k+1,k}^2}}.$$

Left multiplying by $G_k^{(k+1)}$ only affects the last column of $\tilde{Q}_{k-1}^{(k+1)} H_k$, so to update the QR decomposition we need only amend matrix R_{k-1} by a column and a row of zeros. In particular

$$R_k = G_k^{(k+1)} \tilde{Q}_k^{(k+1)} H_k = \begin{bmatrix} & & & \rho_1 \\ & R_{k-1} & & \vdots \\ & & & \rho_{k-1} \\ & & & \sqrt{\rho_k^2 + h_{k+1,k}^2} \\ 0 & \cdots & 0 & 0 \end{bmatrix}.$$

For the right hand side vector of the QR decomposition we have, at step k, that $\boldsymbol{g}^{k-1} = \beta Q_{k-1}^{(k)} \tilde{\boldsymbol{e}}^1$ where $\tilde{\boldsymbol{e}}^1 \in \mathbb{R}^k$. A zero element can be added to the end of vector $\tilde{\boldsymbol{e}}^1$ to obtain $\boldsymbol{e}^1 \in \mathbb{R}^{k+1}$. Then

$$\beta \tilde{Q}_{k-1}^{(k+1)} \boldsymbol{e}^1 = \begin{bmatrix} \boldsymbol{g}^{k-1} \\ 0 \end{bmatrix},$$

with $\boldsymbol{g}^{k-1} = [\gamma_1, \ldots, \gamma_k]^T$. Left-multiplying the above by $G_k^{(k-1)}$ gives us

$$\boldsymbol{g}^k = \beta Q_k^{(k+1)} \boldsymbol{e}^1 = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{k-1} \\ \hat{\gamma}_k \\ \gamma_{k+1} \end{bmatrix},$$

where $\hat{\gamma}_k = c_k \gamma_k$ and $\gamma_{k+1} = -s_k \gamma_k$. This gives a recursive relation for the norm of the residual at step k, namely $\|\boldsymbol{r}^k\|_2 = |\gamma_{k+1}| = |s_k s_{k-1} \cdots s_1 \beta|$. The relation allows us to avoid having to form the solution, $\boldsymbol{y}^k = R_k^{-1} \boldsymbol{g}^k$, to the least squares problem until a specified tolerance in the norm of the residual has been met.

Another issue that arises when implementing the GMRES method is due to the orthonormal basis we construct for the Krylov subspace $\mathcal{K}_k$. Owing to round off errors, as the number of steps, k, of the GMRES algorithm grows, the basis vectors $\{\boldsymbol{v}^i\}$ can become less and less orthogonal. Moreover to perform orthogonalisation the Arnoldi process requires us to store all the previously calculated basis vectors, which increases the memory and computation requirements. A common approach to mitigate both of these issues is to perform a specified number of steps of the GMRES algorithm, say twenty, and then restart the method with approximate solution $\boldsymbol{x}^{20}$ as the new initial guess.

A different approach is to cut down the number of iterations needed in the GMRES algorithm by accelerating it using a *preconditioner*. Let P denote a preconditioner matrix, the idea is to solve the preconditioned linear system

$$P^{-1} A \boldsymbol{x} = P^{-1} \boldsymbol{b}, \tag{4.6}$$

with the GMRES algorithm. The preconditioner P is usually chosen so that it is close to A in some sense but, to be worthwhile, any choice of preconditioner must result in solving

the preconditioned system (4.6) with GMRES being much faster than solving the unpreconditioned system. A preconditioner must be non-singular and inexpensive to implement in the sense that solving the linear system $P\boldsymbol{x} = \boldsymbol{b}$ is not too costly. Algorithm 4.3 shows how the GMRES method is implemented with a preconditioner. The Arnoldi process for

Algorithm 4.3: The preconditioned GMRES method

1 Given tolerance TOL in the residual norm and initial guess $\boldsymbol{x}^0$ compute $\boldsymbol{r}^0 = P^{-1}(b - A\boldsymbol{x}^0)$. Let $\boldsymbol{v}^1 = \boldsymbol{r}^0/\beta$, where $\beta = \|\boldsymbol{r}^0\|_2$.
2 **for** $k = 1, \ldots$ **do**
3 | $\boldsymbol{w}^k = P^{-1}A\boldsymbol{v}^k$.
4 | Lines **4 - 12** are the same as those in Algorithm 4.2.
end

this preconditioned GMRES algorithm will result in a preconditioned Krylov subspace:

$$\mathcal{K}_k = \text{span}\left\{\boldsymbol{r}^0,\ P^{-1}A\boldsymbol{r}^0,\ \ldots,\ (P^{-1}A)^{m-1}\boldsymbol{r}^0\right\}.$$

We discuss 2LM method preconditioners in Chapter 6 when we consider problems with many subdomains and cross points.

4.2 Convergence of the GMRES method

We wish to know how fast GMRES converges when solving the linear system $A\boldsymbol{x} = \boldsymbol{b}$. In particular we are interested in what properties of the system matrix A determine the convergence. A typical stopping criterion for convergence using GMRES is to stop the iteration if the relative residual

$$\frac{\|\boldsymbol{r}^k\|_2}{\|\boldsymbol{r}^0\|_2} \qquad \text{for} \quad k = 1,\ 2,\ \ldots \tag{4.7}$$

reaches a given tolerance, assuming $\|\boldsymbol{r}^0\|_2 \neq 0$.

From the GMRES process described above we know that GMRES converges monotonically, i.e. $\|\boldsymbol{r}^{k+1}\|_2 \leq \|\boldsymbol{r}^k\|_2$. This follows from the fact that at each step of the GMRES iteration $\|\boldsymbol{r}^k\|_2$ is minimised with respect to the Krylov subspace $\mathcal{K}_k$. At the next step of the iteration we enlarge $\mathcal{K}_k$ to $\mathcal{K}_{k+1}$ and minimise $\|\boldsymbol{r}^{k+1}\|_2$ with respect to $\mathcal{K}_{k+1}$. Then $\|\boldsymbol{r}^{k+1}\|_2$ can only be smaller than or at worst equal to $\|\boldsymbol{r}^k\|_2$.

Moreover we know that for an $n \times n$ system $A\boldsymbol{x} = \boldsymbol{b}$, assuming exact arithmetic, the GMRES algorithm converges in at most n steps. However to be of any use, as compared with other methods, we would expect the GMRES algorithm to converge in $m \ll n$ steps.

To gain a more fruitful understanding of its convergence we can reformulate GMRES as a problem of polynomial approximation. Observe that at step k of the GMRES iteration the approximate solution $\boldsymbol{x}^k \in \boldsymbol{x}^0 + \mathcal{K}_k$ can be written using a linear combination of the basis vectors, $\boldsymbol{r}^0$, $A\boldsymbol{r}^0$, $\ldots$, $A^{k-1}\boldsymbol{r}^0$, of the subspace $\mathcal{K}_k$ and the initial guess vector $\boldsymbol{x}^0$:

$$\boldsymbol{x}^k = \boldsymbol{x}^0 + \sum_{i=0}^{k-1} c_i A^i \boldsymbol{r}^0,$$

where $c_i \in \mathbb{C}$. Now we have that

$$\begin{aligned}\boldsymbol{r}^k &= \boldsymbol{b} - A\left(\boldsymbol{x}^0 + \sum_{i=0}^{k-1} c_i A^i \boldsymbol{r}^0\right) \\ &= \left(I - \sum_{i=0}^{k-1} c_i A^{i+1}\right)\boldsymbol{r}^0.\end{aligned}$$

The above can be rewritten as

$$\boldsymbol{r}^k = p_k(A)\boldsymbol{r}^0,$$

where $p_k(z)$ is a polynomial of degree $\leq k$ such that $p_k(0) = 1$. Let $\mathcal{P}_k$ be the set of all polynomials of degree $\leq k$. Then the minimal residual norm property of the GMRES

method can be formulated as the following polynomial approximation problem:

$$\|\boldsymbol{r}^k\|_2 = \min_{\substack{p_k \in \mathcal{P}_k \\ p_k(0)=1}} \|p_k(A)\boldsymbol{r}^0\|_2. \tag{4.8}$$

In words, at step k the GMRES iteration finds a suitable polynomial p_k, normalised to $p_k(0) = 1$, such that the norm of the residual is minimised.

Using (4.8) we get the following bound for (4.7):

$$\frac{\|\boldsymbol{r}^k\|_2}{\|\boldsymbol{r}^0\|_2} \leq \min_{\substack{p_k \in \mathcal{P}_k \\ p_k(0)=1}} |||p_k(A)|||_2, \tag{4.9}$$

where $|||A|||_2 = \max\{\|Ax\|_2 : \|x\|_2 = 1\}$ is the induced matrix 2-norm.

In bound (4.9) we have disentangled the effect of the initial residual on how fast GMRES converges, instead we are interested in minimising the norm of the matrix polynomial. Consider the following inequality:

$$\max_{\boldsymbol{r}^0 \neq 0} \min_{\substack{p_k \in \mathcal{P}_k \\ p_k(0)=1}} \frac{\|p_k(A)\boldsymbol{r}^0\|_2}{\|\boldsymbol{r}^0\|_2} \leq \min_{\substack{p_k \in \mathcal{P}_k \\ p_k(0)=1}} \max_{\boldsymbol{r}^0 \neq 0} \frac{\|p_k(A)\boldsymbol{r}^0\|_2}{\|\boldsymbol{r}^0\|_2} = \min_{\substack{p_k \in \mathcal{P}_k \\ p_k(0)=1}} |||p_k(A)|||_2, \tag{4.10}$$

The right hand side of (4.10) is called the *ideal GMRES* approximation problem and was introduced by Greenbaum and Trefethen, [32], it gives an upper bound for the worst case GMRES approximation given on the left hand side.

How fast a GMRES iteration converges, i.e. how fast $\|p_k(A)\boldsymbol{r}^0\|_2$ converges, depends on both the matrix A and the initial residual $\boldsymbol{r}^0$. In practice however it is the matrix A that will principally determine the convergence. The ideal GMRES approximation lets us disentangle the effect of the initial residual in analysing the convergence of GMRES. The question arises whether removing the effect of the initial residual still gives us a reasonable upper bound on the right hand side of (4.10)?

It is known that for normal matrices the inequality in (4.10) becomes an equality and no problems arise from disentangling the approximation problem from the initial residual, [29, 38]. It was conjectured by Greenbaum and Trefethen that equality held for non-normal matrices as well but counter examples were found in which the right hand side of (4.10) is arbitrarily larger than the left hand side, [18, 62]. In practice, however, such examples are rare, [61], and we are satisfied with the inequality given by (4.9).

Now to understand the convergence of the GMRES method we are interested in what properties of the system matrix A govern how the right hand side of bound (4.9) behaves.

4.2.1 Convergence bound for symmetric indefinite matrices

We first consider GMRES convergence in the case when the system matrix A is symmetric indefinite. In the context of the 2LM method this corresponds to the special situation in which the subdomains are symmetric about the interface.

Theorem 4.2.1. *Let $A \in \mathbb{R}^n$ be symmetric and denote by $\sigma(A) \subset \mathbb{R} \setminus \{0\}$, the spectrum of A and $p(A)$ a polynomial in A. Then*

$$\|\|p(A)\|\|_2 = \max_{\lambda \in \sigma(A)} |p(\lambda)|. \tag{4.11}$$

Proof. Since A is symmetric, there exists a diagonalisation $A = UDU^T$, where $D = diag(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix of the eigenvalues of A and the columns of matrix U are the right eigenvectors of A which form an orthonormal basis for $\mathbb{R}^n$. Using the fact that U is orthogonal, $U^TU = I$, it follows that

$$A^2 = UDU^TUDU^T = UD^2U^T$$

and

$$A^3 = UDU^TUD^2U^T = UD^3U^T.$$

Hence by induction

$$A^k = UD^kU^T.$$

Then any polynomial of A is of the form

$$p(A) = Up(D)U^T$$

Now using the isometric property of orthogonal matrices, namely $|||UX|||_2 = |||X|||_2 = \left|||XU^T|||\right|_2$, for any X, we have

$$\begin{aligned} |||p(A)|||_2 &= \left|||Up(D)U^T|||\right|_2 \\ &= |||p(D)|||_2 \\ &= |||diag(p(\lambda_1), \ldots, p(\lambda_n))|||_2. \end{aligned}$$

The result in (4.11) follows using the fact that for any diagonal matrix D, the matrix 2-norm is given by $|||D|||_2 = \max_{i=1,\ldots,n} |d_i|$, where d_i are the diagonal entries of D. □

The next result shows that it is the eigenvalues alone that govern the convergence of GMRES when the system matrix is symmetric indefinite. Recall that the 2-norm condition number of a matrix A is $\kappa(A) = \|A\|_2\|A^{-1}\|_2$. If the matrix A is normal then $\kappa(A) = |\lambda_{\max}|/|\lambda_{\min}|$, where $\lambda_{\min}$ and $\lambda_{\max}$ are respectively the smallest and largest, in moduli, eigenvalues of A.

Theorem 4.2.2. *Let system $A\boldsymbol{x} = \boldsymbol{b}$ be solved with the GMRES method and assume matrix A is symmetric indefinite. Let $\lambda_{\min}$ denote the eigenvalue of A whose magnitude is smallest and $\lambda_{\max}$ the eigenvalue of A whose magnitude is largest. Then $|\sigma(A)| \subset [|\lambda_{\min}|, |\lambda_{\max}|]$.*

Let $\kappa = |\lambda_{\max}|/|\lambda_{\min}|$. Then at the k^{th} iterate of the GMRES algorithm

$$\frac{\|\boldsymbol{r}^k\|_2}{\|\boldsymbol{r}^0\|_2} \leq 2\left(\frac{\kappa-1}{\kappa+1}\right)^{\lfloor k/2 \rfloor},$$

where $\lfloor k/2 \rfloor$ denotes the integer part of $k/2$. In particular GMRES converges since $\frac{\kappa-1}{\kappa+1} < 1$.

Proof. From (4.9) and (4.11) we have to find a minimising polynomial $p_k(x)$ of degree $\leq k$ such that $p_k(0) = 1$, which will give us a bound for $\|\|p_k(A)\|\|_2$. First consider the *Chebyshev polynomial of the first kind*, $T_m(x)$ of degree m defined on the interval $[-1, 1]$ by

$$T_m(x) = \cos\left(m\cos^{-1}(x)\right) = \cosh\left(m\cosh^{-1}(x)\right),$$

where

$$T_0(x) = 1, \qquad T_1(x) = 1$$

and

$$T_m(x) = 2zT_{m-1}(x) - T_{m-2}(x).$$

The Chebyshev polynomials have the property that $|T_m(x)| \leq 1$ for all $x \in [-1, 1]$ and the nodes of $T_m(x)$ are spaced in such a way as to minimise the errors that arise from Runge's phenomenon, [55].

We can also define the Chebyshev polynomial for the interval $[a, b] \subset (0, \infty)$:

$$T_m(x)^{[a,b]} = T_m(w),$$

where the change of variables is given by

$$w = \frac{z + 1 - 2x/a}{z - 1},$$

with $z = b/a$.

Now consider the following polynomial for bound (4.9):

$$p_k^*(x) = \frac{T_{\lfloor k/2 \rfloor}^{[\lambda_{\min}^2, \lambda_{\max}^2]}(x^2)}{T_{\lfloor k/2 \rfloor}^{[\lambda_{\min}^2, \lambda_{\max}^2]}(0)}. \tag{4.12}$$

To see that this polynomial meets the criteria, observe that since $T_{\lfloor k/2 \rfloor}(x)$ is a polynomial of degree $\lfloor k/2 \rfloor$ in x, then $T_{\lfloor k/2 \rfloor}(x^2)$ is a polynomial of degree $2\lfloor k/2 \rfloor \leq k$ in x. Moreover $p_k^*(0) = 1$, as required.

First consider the numerator of (4.12), we claim that for any $x^2 \in [\lambda_{\min}^2, \lambda_{\max}^2]$, $|T_{\lfloor k/2 \rfloor}^{[\lambda_{\min}^2, \lambda_{\max}^2]}(x^2)| \leq 1$. To see this note that if $x^2 \in [\lambda_{\min}^2, \lambda_{\min}^2]$, then with $z = \lambda_{\max}^2/\lambda_{\min}^2 = \kappa^2$ the change of variables gives us

$$w = \frac{\kappa^2 + 1 - 2x^2/\lambda_{\min}^2}{\kappa^2 - 1} \in [-1, 1].$$

Then $\cos^{-1}(w) \in [0, \pi]$ and $\cos\left(\lfloor k/2 \rfloor \cos^{-1}(w)\right) \in [-1, 1]$, as required.

It follows that

$$\begin{aligned} \min_{\substack{p_k \in \mathcal{P}_k \\ p_k(0)=1}} \|\!|p_k(A)|\!\|_2 = \max_{\lambda \in \sigma(A)} |p_k^*(\lambda)| &= \max_{\lambda \in [\lambda_{\min}^2, \lambda_{\max}^2]} \left| \frac{T_{\lfloor k/2 \rfloor}^{[\lambda_{\min}^2, \lambda_{\max}^2]}(\lambda^2)}{T_{\lfloor k/2 \rfloor}^{[\lambda_{\min}^2, \lambda_{\max}^2]}(0)} \right| \\ &\leq \frac{1}{\cosh\left(\lfloor k/2 \rfloor \cosh^{-1}\left(\frac{\kappa^2+1}{\kappa^2-1}\right)\right)} \end{aligned}$$

and using the result, $\cosh(x) = (e^x + e^{-x})/2 > e^x/2$ we have

$$\begin{aligned} \max_{\lambda \in \sigma(A)} |p_k(\lambda)| &\leq 2 \exp\left(-\lfloor k/2 \rfloor \cosh^{-1}\left(\frac{\kappa^2+1}{\kappa^2-1}\right)\right) \\ &= 2 \exp\left(-\cosh^{-1}\left(\frac{\kappa^2+1}{\kappa^2-1}\right)\right)^{\lfloor k/2 \rfloor}. \end{aligned}$$

The trigonometric identity $\exp(\cosh^{-1}(x)) = x + \sqrt{x+1}\sqrt{x-1}$ gives us the final result:

$$\begin{aligned}\exp\left(-\cosh^{-1}\left(\frac{\kappa^2+1}{\kappa^2-1}\right)\right) &= \left(\frac{\kappa^2+1}{\kappa^2-1} + \sqrt{\frac{\kappa^2+1}{\kappa^2-1}+1}\sqrt{\frac{\kappa^2+1}{\kappa^2-1}-1}\right)^{-1} \\ &= \frac{\kappa-1}{\kappa+1},\end{aligned}$$

as required. □

The convergence bound given in (4.11) tells us that for a symmetric indefinite matrix the smaller the interval $[|\lambda_{\min}|, |\lambda_{\max}|]$ that $|\sigma(A)|$ is contained in, the faster the convergence of GMRES will be. In general, however, the problems we want to solve with the 2LM method will have subdomains that are not symmetric about the interface and as a result the system matrix will be non-symmetric.

4.2.2 Convergence bounds for non-symmetric matrices

Again we seek a bound for the right hand side of (4.9). There are three approaches that depend on the properties of matrix A, [17, 28].

Eigenvalues and eigenvectors

The first approach is to consider an eigen-decomposition as we did in the case of a symmetric indefinite matrix. Assume that matrix A is diagonalisable, then we have $A = V\Lambda V^{-1}$. Here $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix of the complex valued eigenvalues of A and the columns of V are the right eigenvectors of A. Then applying Theorem 4.2.1 to the symmetric matrix Λ we can derive the following bound for (4.9):

$$\frac{\|\boldsymbol{r}^k\|_2}{\|\boldsymbol{r}^0\|_2} \leq \kappa(V) \min_{\substack{p_k \in \mathcal{P}_k \\ p_k(0)=1}} \max_{i=1,\ldots,n} |p_k(\lambda_i)|, \tag{4.13}$$

where $\kappa(V) = \|\|V\|\|_2 \|\|V^{-1}\|\|_2$ is the 2-norm condition number of V.

If A is normal, i.e A is a diagonalisable matrix with a complete set of orthonormal eigenvectors, then $\kappa(V) = 1$ and bound (4.13) is sharp, [29]. Now the problem of determining GMRES convergence reduces to an approximation problem like that in Theorem 4.2.2. In [58] a result analogous to Theorem 4.2.2 using complex valued Chebyshev polynomials is derived for a normal matrix whose eigenvalues are assumed to be contained in an ellipse that does not contain the origin. The bound in (4.13) can also be informative if A is near to normal, so that $\kappa(V) \approx 1$.

If A is highly non-normal and $\kappa(V) \gg 1$ then bound (4.13) may no longer be sharp and studying the eigenvalues alone gives no useful information about the convergence of GMRES, [42, 43]. In particular it was shown in [30] that any non-increasing convergence curve can be obtained for GMRES applied to a non-normal system matrix. Moreover that matrix can be chosen to have any desired eigenvalues.

The resolvent norm

Recognising the deficiencies of studying eigenvalues alone, a second approach to approximating the right hand side of (4.9), given in [51], is to consider the *resolvent norm* $\|\|(zI - A)^{-1}\|\|_2$. For any polynomial $p \in \mathcal{P}_k$, the matrix polynomial $p(A)$ can be written as the following Dunford-Taylor integral, [41]:

$$p(A) = \frac{1}{2\pi i} \int_\gamma p(z)(zI - A)^{-1} dz, \tag{4.14}$$

where γ is a closed curve that contains the spectrum $\sigma(A)$ of matrix A. Applying norms to both sides of (4.14) yields

$$\|\|p(A)\|\|_2 \leq \frac{\mathcal{L}(\gamma)}{2\pi} \max_{z \in \gamma} \left\|\left|p(z)(zI - A)^{-1}\right|\right\|_2, \tag{4.15}$$

where $\mathcal{L}(\gamma)$ is the length of the curve γ. Now fixing $\epsilon > 0$ and considering a curve γ_ϵ on which the resolvent norm $\|\|(zI - A)^{-1}\|\|_2 = \epsilon^{-1}$, (4.15) becomes

$$\|\|p(A)\|\|_2 \leq \frac{\mathcal{L}(\gamma_\epsilon)}{2\pi\epsilon} \max_{z\in\gamma_\epsilon} |p(z)|$$

and we have the following bound for the convergence of GMRES:

$$\frac{\|\boldsymbol{r}^k\|_2}{\|\boldsymbol{r}^0\|_2} \leq \frac{\mathcal{L}(\gamma_\epsilon)}{2\pi\epsilon} \min_{\substack{p_k\in\mathcal{P}_k \\ p_k(0)=1}} \max_{z\in\gamma_\epsilon} |p_k(z)|. \tag{4.16}$$

The bound above can also be derived with reference to the *pseudospectra* of matrix A, [64]. This is because the curve γ_ϵ is the boundary of the ϵ-pseudospectrum of A defined as:

$$\Lambda_\epsilon(A) = \{z \in \mathbb{C} : \left\|\left|(zI - A)^{-1}\right|\right\|_2 \geq \epsilon^{-1}\}.$$

By choosing different values of ϵ bound (4.16) changes and this can be useful as different values of ϵ may describe the convergence of GMRES at different stages of the iteration [17]. The difficulty arises in that calculating the resolvent norm may be expensive and choosing the correct ϵ is tricky. There is a trade off in the two terms of bound (4.16). We must choose an ϵ large enough such that the term $\frac{\mathcal{L}(\gamma_\epsilon)}{2\pi\epsilon}$ is small but not too large so as to leave the curve γ_ϵ encasing the spectrum too large. In fact examples have been found where no choice of ϵ gives a reasonable right hand side for bound (4.16), [31].

The field of values

The third approach for studying the convergence of the GMRES method, and the one we focus on in this thesis, is to consider the *field of values* of the system matrix. For an arbitrary matrix $A \in \mathbb{C}^{n\times n}$, the field of values (also known as the numerical range) of

matrix A is the subset of $\mathbb{C}$ defined by

$$W(A) = \{\boldsymbol{w}^* A\boldsymbol{w} : \boldsymbol{w} \in \mathbb{C}^n, \boldsymbol{w}^*\boldsymbol{w} = 1\}.$$

In particular $W(A)$ is the set of all the Rayleigh quotients of matrix A. We now quote some well known results about the field of values that can be found in [35] and [36].

Theorem 4.2.3. *Let A, $B \in \mathbb{C}^{n\times n}$, $\alpha \in \mathbb{C}$ and for sets S, $T \subset \mathbb{C}$ let $S + T = \{s + t : s \in S, t \in T\}$, then the field of values has the following properties:*

(i) (Compactness) $W(A)$ is a compact subset of $\mathbb{C}$.

(ii) (Translation)

$$W(A + \alpha I) = W(A) + \alpha. \tag{4.17}$$

(iii) (Scalar multiplication)

$$W(\alpha A) = \alpha W(A). \tag{4.18}$$

(iv) (Subadditivity)

$$W(A + B) \subset W(A) + W(B).$$

(v) (Spectral containment)

$$\sigma(A) \subset W(A).$$

(vi) (Unitary similarity invariance) If $U \in \mathbb{C}^{n\times n}$ is unitary then

$$W(U^* AU) = W(A). \tag{4.19}$$

(vii) (Normality) If A is normal then $W(A) = Co(\sigma(A))$, where $Co(X)$ denotes the convex hull of the set X, the smallest convex set that contains X.

Proof. (i) The set $W(A)$ is the range of the continuous function $\boldsymbol{w} \mapsto \boldsymbol{w}^* A\boldsymbol{w}$ over the domain $\{\boldsymbol{w} : \boldsymbol{w} \in \mathbb{C}^n, \boldsymbol{w}^*\boldsymbol{w} = 1\}$, the surface of the unit ball, which is a compact set. Now, since the continuous image of a compact set is compact itself, $W(A)$ is compact.

(ii)

$$\begin{aligned} W(A + \alpha I) &= \{\boldsymbol{w}^*(A + \alpha I)\boldsymbol{w} : \boldsymbol{w}^*\boldsymbol{w} = 1\} \\ &= \{\boldsymbol{w}^* A\boldsymbol{w} + \alpha\boldsymbol{w}^*\boldsymbol{w} : \boldsymbol{w}^*\boldsymbol{w} = 1\} \\ &= \{\boldsymbol{w}^* A\boldsymbol{w} : \boldsymbol{w}^*\boldsymbol{w} = 1\} + \alpha \\ &= W(A) + \alpha. \end{aligned}$$

(iii)

$$\begin{aligned} W(\alpha A) &= \{\boldsymbol{w}^*(\alpha A)\boldsymbol{w} : \boldsymbol{w}^*\boldsymbol{w} = 1\} \\ &= \{\alpha\boldsymbol{w}^* A\boldsymbol{w} : \boldsymbol{w}^*\boldsymbol{w} = 1\} \\ &= \alpha W(A). \end{aligned}$$

(iv)

$$\begin{aligned} W(A + B) &= \{\boldsymbol{w}^*(A + B)\boldsymbol{w} : \boldsymbol{w} \in \mathbb{C}^n, \boldsymbol{w}^*\boldsymbol{w} = 1\} \\ &= \{\boldsymbol{w}^* A\boldsymbol{w} + \boldsymbol{w}^* B\boldsymbol{w} : \boldsymbol{w} \in \mathbb{C}^n, \boldsymbol{w}^*\boldsymbol{w} = 1\} \\ &\subset \{\boldsymbol{w}^* A\boldsymbol{w} : \boldsymbol{w} \in \mathbb{C}^n, \boldsymbol{w}^*\boldsymbol{w} = 1\} + \{\boldsymbol{z}^* B\boldsymbol{z} : \boldsymbol{z} \in \mathbb{C}^n, \boldsymbol{z}^*\boldsymbol{z} = 1\} \\ &= W(A) + W(B). \end{aligned}$$

(v) Consider $\lambda \in \sigma(A)$. Then there exists a non-zero $\boldsymbol{w} \in \mathbb{C}^n$, with $\boldsymbol{w}^*\boldsymbol{w} = 1$, for which

$A\boldsymbol{w} = \lambda\boldsymbol{w}$, then

$$\lambda = \lambda\boldsymbol{w}^*\boldsymbol{w} = \boldsymbol{w}^*(\lambda\boldsymbol{w}) = \boldsymbol{w}^*A\boldsymbol{w} \in W(A).$$

(vi) Since unitary transformations leave the surface of the unit ball unchanged, $W(A)$ and $W(U^*AU)$ are comprised of the same complex numbers. Let $\boldsymbol{w} \in \mathbb{C}^n$ with $\boldsymbol{w}^*\boldsymbol{w} = 1$, we have that $\boldsymbol{w}^*(U^*AU)\boldsymbol{w} = \boldsymbol{z}^*A\boldsymbol{z}$, where $\boldsymbol{z} = U\boldsymbol{w}$, so

$$\boldsymbol{z}^*\boldsymbol{z} = \boldsymbol{w}^*U^*U\boldsymbol{w} = \boldsymbol{w}^*\boldsymbol{w} = 1.$$

Hence $W(U^*AU) \subset W(A)$. The reverse result, $W(A) \subset W(U^*AU)$, follows in a similar manner.

(vii) If A is normal there exists a decomposition such that $A = U^*DU$, where $D = diag(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix whose entries are the eigenvalues of A and U is a unitary matrix. Then from the unitary invariance result (vi) we have that $W(A) = W(D)$ and since, for $\boldsymbol{w} = (w_1, \ldots, w_n)^T$,

$$\boldsymbol{w}^*D\boldsymbol{w} = \sum_{i=1}^{n} \bar{w}_i w_i \lambda_i = \sum_{i=1}^{n} |w_i|^2 \lambda_i,$$

$W(D)$ is the set of all convex combinations of the diagonal entries of D ($\boldsymbol{w}^*\boldsymbol{w} = 1$ implies $\sum_i^n |w_i|^2 = 1$ and $|w_i|^2 \geq 0$). Now since the diagonal entries of D are the eigenvalues of A, it follows that $W(A) = Co(\sigma(A))$.

□

It follows, as a consequence of property (vii), that if matrix A is Hermitian the field of values of A is an interval on the real line with endpoints given by its smallest and largest eigenvalues. The general result that for any matrix A, its field of values is a convex set is known as the Toeplitz-Hausdorff Theorem, [35].

Associated with the field of values is the *numerical radius*, defined by

$$r(A) = \max\{|z| : z \in W(A)\},$$

i.e. the largest absolute value of an element of $W(A)$.

Theorem 4.2.4. *Let A, $B \in \mathbb{C}^{n\times n}$, then the numerical radius has the following properties:*

(i) (Conjugate transpose)

$$r(A^*) = r(A). \tag{4.20}$$

(ii) (Subadditivity)

$$r(A+B) \leq r(A) + r(B). \tag{4.21}$$

(iii) (Normality) Let $\lambda_{\max}$ be the eigenvalue of A whose magnitude is largest. If A is normal then

$$r(A) = |\lambda_{\max}| = |||A|||_2. \tag{4.22}$$

(iv) Though $r(A)$ itself is not a matrix norm (since the requirement that $r(AB) \leq r(A)r(B)$ does not hold for all A and B), the numerical radius is connected to the matrix 2-norm through the relation:

$$|||A|||_2 \leq 2r(A) \leq 2|||A|||_2. \tag{4.23}$$

Proof. (i)

$$\begin{aligned} r(A) &= \max_{\boldsymbol{w}^*\boldsymbol{w}=1} |\boldsymbol{w}^* A \boldsymbol{w}| \\ &= \max_{\boldsymbol{w}^*\boldsymbol{w}=1} |(A^*\boldsymbol{w})^*\boldsymbol{w}| \\ &= \max_{\boldsymbol{w}^*\boldsymbol{w}=1} |(\boldsymbol{w}^*(A^*\boldsymbol{w}))^*| \\ &= r(A^*). \end{aligned}$$

(ii)

$$\begin{aligned} r(A+B) &= \max_{\boldsymbol{w}^*\boldsymbol{w}=1} |\boldsymbol{w}^*(A+B)\boldsymbol{w}| \\ &\leq \max_{\boldsymbol{w}^*\boldsymbol{w}=1} |\boldsymbol{w}^*A\boldsymbol{w}| + \max_{\boldsymbol{w}^*\boldsymbol{w}=1} |\boldsymbol{w}^*B\boldsymbol{w}| \\ &= r(A) + r(B). \end{aligned}$$

(iii) Since A is normal there exists a decomposition $A = U^*DU$, where $D = diag(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix with whose entries are the eigenvalues of A and U is unitary matrix. Then using (4.19) we have that

$$r(A) = r(U^*DU) = \max_{z \in W(U^*DU)} |z| = \max_{z \in W(D)} |z| = r(D).$$

Now since D is a diagonal matrix its field of values is an interval on the real line and $r(D) = \max_{i=1,\ldots,n} |\lambda_i| = |||A|||_2$, where the final result follows from the fact that for a normal matrix its 2-norm is its spectral radius.

(iv) First for the inequality on the right hand side of (4.23) we have that

$$r(A) = \max_{\boldsymbol{w}^*\boldsymbol{w}=1} |\boldsymbol{w}^*A\boldsymbol{w}| \leq \max_{\boldsymbol{w}^*\boldsymbol{w}=1} \|\boldsymbol{w}^*\|_2 \|\boldsymbol{w}\|_2 |||A|||_2 = |||A|||_2.$$

For the inequality on the left hand side, let

$$A_1 = \frac{A + A^*}{2} \qquad \text{and} \qquad A_2 = \frac{A - A^*}{2},$$

then $A = A_1 + A_2$. Now since both A_1 and A_2 are normal it follows from (4.22) that

$$\begin{aligned} |||A|||_2 &\leq \frac{1}{2}|||A + A^*|||_2 + \frac{1}{2}|||A - A^*|||_2 \\ &= \frac{1}{2}r(A + A^*) + \frac{1}{2}r(A - A^*) \\ &\leq \frac{1}{2}\left(2r(A) + 2r(A^*)\right) \\ &= 2r(A). \end{aligned}$$

□

The numerical radius also satisfies the power inequality:

$$r(A^m) \leq [r(A)]^m \quad \text{for} \quad m = 1, 2, \ldots \tag{4.24}$$

See [53] for a proof.

The following two results give bounds for GMRES convergence, if the field of values of the system matrix is contained inside a disk or an ellipse.

Theorem 4.2.5. *If $0 \notin W(A)$ and $W(A)$ is contained in the disk given by $\mathcal{D} = \{z \in \mathbb{C} : |z - c| \leq s\}$ which does not contain the origin. Then the relative residual for the GMRES method satisfies*

$$\frac{\|\boldsymbol{r}^k\|_2}{\|\boldsymbol{r}^0\|_2} \leq 2\left(\frac{s}{|c|}\right)^k.$$

Proof. Consider the polynomial $p_k^* = (1 - z/c)^k$. From (4.17) and (4.18) we have that

$$W(I - (1/c)A) = 1 - (1/c)W(A) \subset \{z \in \mathbb{C} : |z| \leq s/|c|\}.$$

It follows that the numerical radius satisfies $r(I-(1/c)A) \leq s/|c|$. Now the power inequality

gives us

$$r(p_k^*(A)) = r((I-(1/c)A)^k) \leq (r(I-(1/c)A))^k \leq \left(\frac{s}{|c|}\right)^k.$$

The result follows from (4.9) and (4.23):

$$\frac{\|\boldsymbol{r}^k\|_2}{\|\boldsymbol{r}^0\|_2} \leq \|\!|p_k^*(A)|\!\|_2 \leq 2r(p_k^*(A)) \leq 2\left(\frac{s}{|c|}\right)^k.$$

□

The deficiency in the above estimate is that the condition that $0 \notin W(A)$ may leave the disk needed to contain $W(A)$ being very large and the bound for the convergence unreliable. We can improve the estimate if we can bound the field of values by an ellipse. The following result is due to Eiermann, which we quote from [15].

Theorem 4.2.6. *Let*

$$\kappa = \left|\frac{\delta - \sqrt{\delta^2 - \tau^2}}{\tau^2}\right|$$

where the branch of the square root is chosen such that $\kappa < 1$. If $0 \notin W(A)$ and $W(A)$ is contained in the ellipse, which does not contain the origin, given by

$$\mathcal{E}_s = \{z \in \mathbb{C} : |z-(\delta-\tau)| + |z-(\delta+\tau)| \leq |\tau|(s+s^{-1})\}$$

with foci $\delta \pm \tau$, semi-axes $\tau(s+s^{-1})$ and where $s < \kappa^{-1}$. Then the relative residual of the GMRES method satisfies

$$\frac{\|\boldsymbol{r}^k\|_2}{\|\boldsymbol{r}^0\|_2} \leq 2(s^k + s^{-k})\frac{\kappa^k}{1-\kappa^{2k}}.$$

Proof. Consider the polynomial

$$p_k^*(z) = \frac{T_k^{[\delta-\tau,\delta+\tau]}(z)}{T_k^{[\delta-\tau,\delta+\tau]}(0)}.$$

The image $p_k^*(\partial\mathcal{E}_s)$ of $\partial\mathcal{E}_s$ under p_k^* is $\partial\tilde{\mathcal{E}}_{s^k}$ (covered exactly k times), where

$$\tilde{\mathcal{E}}_{s^k} = \frac{1}{T_k^{[\delta-\tau,\delta+\tau]}(0)}\{z \in \mathbb{C} : \ |z+1| + |z-1| \leq s^k + s^{-k}\},$$

i.e. $(p_k^*)^{-1}(\tilde{\mathcal{E}}_{s^k}) = \mathcal{E}_s$. Now, since both $\mathcal{E}_s$ and $\tilde{\mathcal{E}}_{s^k}$ are compact and convex, it follows from a mapping theorem for the numerical radius, [40], that $W(p_k^*(A)) \subset \tilde{\mathcal{E}}_{s^k}$. Hence $r(p_k^*(A)) \leq \max\{|p_k^*(z)| : \ z \in \mathcal{E}_s\}$ with (4.9) and (4.23) giving us

$$\frac{\|\boldsymbol{r}^k\|_2}{\|\boldsymbol{r}^0\|_2} \leq |||p_k^*(A)|||_2 \leq 2r(p_k^*(A)) \leq 2\max_{z\in\mathcal{E}_s}|p_k^*(z)|.$$

The final result follows using known properties of Chebyshev polynomials, [14]. □

Another estimate for GMRES convergence is given by Elman, [16], which says that if $0 \notin W(A)$ we have that $dist(0, W(A)) = \lambda_{\min}((A + A^T)/2)$ and for $\beta \in (0, \pi/2)$

$$\frac{\|\boldsymbol{r}^k\|_2}{\|\boldsymbol{r}^0\|_2} \leq \sin^k(\beta),$$

with

$$\cos(\beta) = \frac{dist(0, W(A))}{|||A|||_2}.$$

Beckermann et al, [2], have improved on the Elman estimate to give

$$\frac{\|\boldsymbol{r}^k\|_2}{\|\boldsymbol{r}^0\|_2} \leq (2 + 2\sqrt{3})(2 + \gamma_\beta)\gamma_\beta^k,$$

where

$$\gamma_\beta = 2\sin\left(\frac{\beta}{4 - 2\beta/\pi}\right) < \sin(\beta).$$

A recent result giving a bound for $\|\|\|p_k(A)\|\|\|_2$ in terms of the field of values is due to Crouzeix. For 2×2 matrices Crouzeix derived the following bound

$$\|\|\|p_k(A)\|\|\|_2 \leq 2 \max_{z \in W(A)} |p_k(z)|, \tag{4.25}$$

[5]. While for general $n \times n$ matrices it was proved that

$$\|\|\|p_k(A)\|\|\|_2 \leq 11.08 \max_{z \in W(A)} |p_k(z)|,$$

[6]. In the same paper Crouzeix conjectured that in fact the stronger bound (4.25) also holds for all $n \times n$ matrices. So far *Crouzeix's conjecture* has not been disproved either theoretically or through numerical experiments.

4.2.3 The asymptotic convergence factor

For each of the three bounds (4.13), (4.16) and (4.25) there is associated a constant term

$$C_{\sigma(A)} = \kappa(V), \qquad C_{\gamma_\epsilon} = \frac{\mathcal{L}(\gamma_\epsilon)}{2\pi\epsilon}, \quad \text{and} \quad C_{W(A)} = 2. \tag{4.26}$$

Assuming the constant terms in the eigenvalue and resolvent norm bounds are not too large, for all three bounds the convergence of the GMRES method is determined by the polynomial approximation problem

$$\min_{\substack{p_k \in \mathcal{P}_k \\ p_k(0)=1}} \max_{z \in \Sigma} |p_k(z)|, \tag{4.27}$$

where Σ is one of the sets, containing the spectrum, given by the eigenvalues, resolvent norm or field of values. As we have discussed solving the polynomial approximation problem defined by (4.27) can be tricky except in specific cases such as when the field of values is contained in a disk or ellipse. However the problem given in (4.27) becomes easier to handle when we take the limit as $k \to \infty$.

Instead of Σ being one of the sets defined by the eigenvalues, resolvent norm or field of values now let Σ be a compact subset of $\mathbb{C}$, without isolated points, that tightly bounds one of these three sets. Because of the normalisation $p_k(0) = 1$, Σ must not contain the origin. Now if Σ does not surround the origin, in the sense that it separates 0 and ∞ in the extended complex plane $\mathbb{C} \cup \{\infty\}$, it is known, [34], that for $k = 1, 2, \ldots$ the sequence defined by (4.27) decreases geometrically with k at some rate

$$\rho = \lim_{k \to \infty} \left(\min_{\substack{p_k \in \mathcal{P}_k \\ p_k(0)=1}} \max_{z \in \Sigma} |p_k(z)| \right)^{1/k} < 1. \tag{4.28}$$

The value ρ is called the *estimated asymptotic convergence factor* for Σ. Then a reasonable estimate for the convergence of the GMRES method is given by

$$\frac{\|\boldsymbol{r}^k\|_2}{\|\boldsymbol{r}^0\|_2} \approx C\rho^k, \tag{4.29}$$

where C is one of the constant terms defined in (4.26) corresponding to whichever set Σ is chosen to bound.

If we are not too rigorous we can derive ρ using potential theory. We follow the procedure as laid out in [8]. Consider the polynomial

$$|p(z)| = \prod_{k=1}^{n} (z - z_k).$$

Then we wish to minimise $|p(z)|/|p(0)|$ on the compact set without isolated points, Σ. By the minimum modulus principle this is equivalent to finding the minimum on $\partial\Sigma$. Taking the logarithm we have

$$\log|p(z)| = \sum_{k=1}^{n} \log|z - z_k|$$

and we wish to minimise

$$\log|p(z)| - \log|p(0)| = \log \prod_{k=1}^{n} \left|1 - \frac{z}{z_k}\right|,$$

on $\partial\Sigma$. Physically this function can be thought of as the potential on $\mathbb{C}$ with electric charges of amplitude -1 at the points $\{z_k\}$.

Minimisation of the above function is difficult but can be simplified by taking the limit $n \to \infty$. First by rescaling the problem we have

$$g(z) = n^{-1} \sum_{k=1}^{n} \log|z - z_k| + C. \tag{4.30}$$

Then the above function is the potential of point charges of amplitude $-n^{-1}$ at each point $\{z_k\}$. Taking the limit $n \to \infty$ we can imagine a negative unit charge distributed in a continuous fashion in $\mathbb{C}$ and we wish to minimise $\max g(z) - g(0)$ on Σ. This minimum will be achieved when $g(z)$ is constant in $\partial\Sigma$.

Physically we can think of Σ as a collection of conductors in $\mathbb{C}$ that are connected. By inserting a charge of -1 into the system we allow an equilibrium to be achieved. The charge will distribute along $\partial\Sigma$ such that the potential $g(z)$ is constant on $\partial\Sigma$. We add a constant C to the potential $g(z)$ so that this constant value becomes 0. Now the asymptotic convergence factor is

$$\rho = \exp(-g(0)). \tag{4.31}$$

Mathematically the potential $g(z)$ is the Green's function associated with Σ. Recall

the Green's function $g(z)$ for a domain Σ is the unique function defined in the exterior of Σ such that $\nabla g = 0$ outside Σ, $g(z) \to 0$ as $z \to \partial\Sigma$ and $g(z) - \log|z| \to C$ as $|z| \to \infty$ for some constant C.

In the case that Σ is connected the problem of finding ρ simplifies. In this case the potential $g(z)$ can be seen as a level curve function of a conformal map. Recall that an analytic function $f(z)$ on a domain $G \subset \mathbb{C}$ is a *conformal mapping* at every point $z \in G$ where $f'(z) \neq 0$. In particular a conformal map preserves angles between curves passing through the same point.

Since Σ is simply connected its exterior is simply connected with respect to the extended complex plane $\mathbb{C} \cup \{\infty\}$. Then there exists a harmonic function $h(z)$ in the exterior of Σ that is the harmonic conjugate of $g(z)$. Now consider the function

$$\Phi(z) = \frac{1}{\exp(g(z) + ih(z))}, \tag{4.32}$$

defined in the exterior of Σ. Then $\Phi(z)$ is a conformal map from the exterior of Σ to the interior of the unit disk, with $\Phi(\infty) = 0$, unique up to an arbitrary rotation.

By assumption the origin is exterior to Σ then its image $\Phi(0)$ is interior to the unit disk. Combining (4.31) and (4.32) we have a formula for ρ.

Theorem 4.2.7. *Let Σ be a simply connected set in the complex plane and let $\Phi(z)$ be the conformal map from the exterior of Σ to the interior of the unit disk such that $\Phi(\infty) = 0$. The estimated asymptotic convergence factor of Σ is*

$$\rho = |\Phi(0)|. \tag{4.33}$$

A full mathematical derivation for both of the formulas for ρ is given in [52]. We give two examples of some conformal maps for elementary sets in $\mathbb{C}$ and the convergence factors

that they produce.

First consider the disk $D = \{|z - z_c| \leq r : z \in \mathbb{C}\}$ for some $r < |z_c|$. The conformal map from the exterior of D to the interior of the unit disk is given by $\Phi(z) = r/|z - z_c|$ and so $\rho = r/z_c$. Then the smaller the disk is the smaller the convergence factor ρ will be. We can redefine the disk D in terms of its "condition number", κ, the ratio of its largest to smallest points. This will allow us to compare its convergence factor with our second example. For our disk D we have $z_c = \beta(\kappa + 1)/2$ and $r = |\beta|(\kappa - 1)/2$ for some constant β. Now the estimated asymptotic convergence factor for the disk is

$$\rho = \frac{\kappa - 1}{\kappa + 1}. \tag{4.34}$$

Now consider the interval $J = [1, \kappa]$. The conformal map from the exterior of J to the interior of the unit disk is given by $\Phi(z) = (\kappa - 1)/(2z - \kappa - 1 + 2\sqrt{z^2 - (\kappa + 1)z + \kappa})$. The estimated asymptotic convergence factor for the interval is

$$\rho = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}. \tag{4.35}$$

We see from the above that the smaller the interval J is the smaller ρ will be.

Comparing (4.34) and (4.35) we see that due to the square root terms the convergence factor is smaller in the example with the interval. This tells us that the convergence of the GMRES method will be faster when applied to a system matrix whose spectrum can be bounded by a small set away from the origin. Furthermore convergence is faster if the spectrum is on the real line.

Following the results of the chapter we will proceed as follows. We will bound the field of values of the 2LM method system matrix by a rectangle. The size of the rectangle will depend on the Robin parameters so by choosing suitable parameters we wish to make the rectangle as small and as far away from the origin as possible. Next we find a conformal

map from the exterior of the rectangle to the interior of the unit disk with which we can compute the estimated asymptotic convergence factor for the GMRES method. Then we are interested in the asymptotic convergence behaviour as the discretisation of the problem becomes small ($h \to 0$) and when the jump in diffusion coefficients becomes large ($\alpha_2 \to 0$, for a fixed α_1).

Chapter 5

Optimised Robin parameters for the 2LM method

5.1 Approximation of the field of values of the 2LM system matrix by a rectangle

A simple numerical algorithm due to Johnson, [37], can be used to approximate the field of values of a matrix $A \in \mathbb{C}^{n\times n}$. It uses the property that any matrix can be split into Hermitian and skew-Hermitian parts. Let $H(A) = \frac{1}{2}(A + A^*)$ denote the Hermitian part of matrix A while $\sigma_{\min}(B)$ and $\sigma_{\max}(B)$ denote the smallest and largest eigenvalues of a Hermitian matrix B respectively. Now, using the fact that the field of values of a Hermitian matrix is an interval on the real line, we have that

$$\min_{z\in W(A)} \Re e(z) = \min_{\gamma\in W(H(A))} \gamma = \sigma_{\min}(H(A))$$

and

$$\max_{z\in W(A)} \Re e(z) = \max_{\gamma\in W(H(A))} \gamma = \sigma_{\max}(H(A))$$

Then $W(A)$ lies in between the lines that run parallel to the imaginary axis, cross the real axis at points $\sigma_{\min}(H(A))$ and $\sigma_{\max}(H(A))$ and which intersect $W(A)$ on its boundary. Now, using the property that $W(e^{i\varphi}A) = e^{i\varphi}W(A)$, we can calculate $\sigma_{\min}(H(e^{i\varphi}A))$ and $\sigma_{\max}(H(e^{i\varphi}A))$ for different angles φ to find boundary points of $W(A)$. If A is real $W(A)$ is symmetric with respect to the real axis and we need only take $\varphi \in [0, \pi/2]$. Using this approach we can approximate the field of values of the 2LM method system matrix by a rectangle in $\mathbb{C}$. We first give a result for general matrices that have a special block structure.

Lemma 5.1.1. *Let X and Y be real, symmetric matrices of the same size and let $\tau \in \mathbb{R}$. Consider matrix A of the form:*

$$A = \begin{bmatrix} \tau I & X \\ Y & \tau I \end{bmatrix}.$$

Then $W(A)$ is contained in a rectangle in $\mathbb{C}$ defined by

$$\left\{ x + iy : |x - \tau| < \frac{1}{2}\rho(X + Y) \quad \textit{and} \quad |y| < \frac{1}{2}\rho(X - Y), \quad x, y \in \mathbb{R} \right\},$$

where $\rho(\cdot)$ denotes the spectral radius of a symmetric matrix.

Proof. For $\varphi \in [0, \pi/2]$ we have that

$$\begin{aligned} H(e^{i\varphi}A) &= \frac{e^{i\varphi}}{2}\begin{bmatrix} \tau I & X \\ Y & \tau I \end{bmatrix} + \frac{e^{-i\varphi}}{2}\begin{bmatrix} \tau I & Y \\ X & \tau I \end{bmatrix} \\ &= \begin{bmatrix} \tau\cos\varphi I & \frac{1}{2}(e^{i\varphi}X + e^{-i\varphi}Y) \\ \frac{1}{2}(e^{i\varphi}Y + e^{-i\varphi}X) & \tau\cos\varphi I \end{bmatrix}. \end{aligned}$$

Assume that A is $n \times n$, let $\boldsymbol{w} \in \mathbb{C}^n$ be such that $\boldsymbol{w}^*\boldsymbol{w} = 1$ and partition $\boldsymbol{w}$ into block form:

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix},$$

so $\boldsymbol{u}^*\boldsymbol{u}+\boldsymbol{v}^*\boldsymbol{v}=1$. Let $z_\varphi = \boldsymbol{w}^* H(e^{i\varphi}A)\boldsymbol{w}$, then $\{e^{-i\varphi}(z_\varphi+\xi i) : \xi \in \mathbb{R}\}$ defines a supporting hyperplane for $W(A)$, where

$$\begin{aligned} z_\varphi &= \begin{bmatrix} \boldsymbol{u}^* & \boldsymbol{v}^* \end{bmatrix} \begin{bmatrix} \tau\cos\varphi I & \frac{1}{2}(e^{i\varphi}X + e^{-i\varphi}Y) \\ \frac{1}{2}(e^{i\varphi}Y + e^{-i\varphi}X) & \tau\cos\varphi I \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} \\ &= \tau\cos\varphi + \frac{1}{2}\boldsymbol{u}^*\left(e^{i\varphi}X + e^{-i\varphi}Y\right)\boldsymbol{v} + \frac{1}{2}\boldsymbol{v}^*\left(e^{i\varphi}Y + e^{-i\varphi}X\right)\boldsymbol{u}. \end{aligned}$$

Taking the absolute value and using Young's inequality, $ab \le \frac{a^2}{2} + \frac{b^2}{2}$, gives

$$\begin{aligned} |z_\varphi - \tau\cos\varphi| &\le \frac{1}{2}\left|\boldsymbol{u}^*(e^{i\varphi}X + e^{-i\varphi}Y)\boldsymbol{v}\right| + \frac{1}{2}\left|\boldsymbol{v}^*(e^{i\varphi}Y + e^{-i\varphi}X)\boldsymbol{u}\right| \\ &\le \frac{1}{2}||\boldsymbol{u}||_2||\boldsymbol{v}||_2\left(\left|\left|\left|e^{i\varphi}X + e^{-i\varphi}Y\right|\right|\right|_2 + \left|\left|\left|e^{i\varphi}Y + e^{-i\varphi}X\right|\right|\right|_2\right) \\ &\le \frac{1}{2}\left(\frac{||\boldsymbol{u}||_2^2}{2} + \frac{||\boldsymbol{v}||_2^2}{2}\right)\left(\left|\left|\left|e^{i\varphi}X + e^{-i\varphi}Y\right|\right|\right|_2 + \left|\left|\left|e^{i\varphi}Y + e^{-i\varphi}X\right|\right|\right|_2\right) \\ &= \frac{1}{4}\left(\left|\left|\left|e^{i\varphi}X + e^{-i\varphi}Y\right|\right|\right|_2 + \left|\left|\left|e^{i\varphi}Y + e^{-i\varphi}X\right|\right|\right|_2\right). \end{aligned} \tag{5.1}$$

These matrix norms are in general difficult to estimate, but we have two special values of φ that simplify bound (5.1). When $\varphi = 0$:

$$\begin{aligned} |z_0 - \tau| &\le \frac{1}{4}|||X+Y|||_2 + \frac{1}{4}|||Y+X|||_2 \\ &= \frac{1}{2}\rho(X+Y), \end{aligned} \tag{5.2}$$

where we have used the fact that $|||A|||_2 = \rho(A)$, if A is symmetric. While for $\varphi = \pi/2$:

$$\begin{aligned} \left|z_{\frac{\pi}{2}}\right| &\le \frac{1}{4}|||iX - iY|||_2 + \frac{1}{4}|||iY - iX|||_2 \\ &= \frac{1}{2}\rho(X-Y). \end{aligned} \tag{5.3}$$

Then the lines that run through the points defined by (5.2) and (5.3) form a rectangle in $\mathbb{C}$ centred at the point $(\tau, 0)$, with top and bottom parallel to the real axis and sides parallel to the imaginary axis, that contains $W(A)$. □

Corollary 5.1.2. *Let* $X = \frac{1}{2}I - p_s\left(\alpha_2 S_2 + p_2 hI\right)^{-1}$, $Y = \frac{1}{2}I - p_s\left(\alpha_1 S_1 + p_1 hI\right)^{-1}$ *and* $\tau = 1/2$. *The 2LM method system matrix is of the form:*

$$A_{2LM} = \begin{bmatrix} \frac{1}{2}I & X \\ Y & \frac{1}{2}I \end{bmatrix}.$$

Then $W(A_{2LM})$ *is contained in the rectangle*

$$\mathbf{R} = \left\{ x + iy : |x - \frac{1}{2}| < \frac{1}{2}\mathcal{R}(p_1, p_2) \quad \textit{and} \quad |y| < \frac{1}{2}\mathcal{I}(p_1, p_2), \quad x, y \in \mathbb{R} \right\},$$

here

$$\mathcal{R}(p_1, p_2) = \max\left\{ \left|\mu_1(p_1, p_2)\right|, \left|\mu_2(p_1, p_2)\right| \right\},$$

where

$$\mu_1(p_1, p_2) = 1 - p_s\left(\frac{1}{\alpha_1 s_{\min} + p_1 h} + \frac{1}{C_1 \alpha_2 s_{\min} + p_2 h} \right)$$

and

$$\mu_2(p_1, p_2) = 1 - p_s\left(\frac{1}{\alpha_1 s_{\max} + p_1 h} + \frac{1}{C_2 \alpha_2 s_{\max} + p_2 h} \right),$$

while

$$\mathcal{I}(p_1, p_2) = \max\left\{ \left|\nu_1(p_1, p_2)\right|, \left|\nu_2(p_1, p_2)\right| \right\},$$

with

$$\nu_1(p_1, p_2) = p_s\left(\frac{1}{C_2 \alpha_2 s_{\max} + p_2 h} - \frac{1}{\alpha_1 s_{\min} + p_1 h} \right)$$

and

$$\nu_2(p_1, p_2) = p_s\left(\frac{1}{C_1 \alpha_2 s_{\min} + p_2 h} - \frac{1}{\alpha_1 s_{\max} + p_1 h} \right).$$

Here $s_{\min}$ and $s_{\max}$ denote the smallest and largest eigenvalues of matrix S_1 respectively, while C_1 and C_2 are positive constants independent of h that follow from the spectral equivalence of S_1 and S_2.

Proof. Following Lemma 5.1.1 we find upper bounds for $\rho(X+Y)$ and $\rho(X-Y)$. Let $t_{\min}$ and $t_{\max}$ denote the smallest and largest eigenvalues of S_2 respectively. It is known (see [56] Proposition 4.1.2) that S_1 and S_2 are spectrally equivalent, so there exists positive constants C_1 and C_2 independent of h such that $C_1\boldsymbol{x}^*S_1\boldsymbol{x} \leq \boldsymbol{x}^*S_2\boldsymbol{x} \leq C_2\boldsymbol{x}^*S_1\boldsymbol{x}$, for all $\boldsymbol{x} \in \mathbb{C}^n$. Hence, since both S_1 and S_2 are symmetric positive definite, from the Rayleigh quotients we have that

$$t_{\min} = \min_{\boldsymbol{x}^*\boldsymbol{x}=1} \boldsymbol{x}^*S_2\boldsymbol{x} \geq C_1 \min_{\boldsymbol{x}^*\boldsymbol{x}=1} \boldsymbol{x}^*S_1\boldsymbol{x} = C_1 s_{\min} \tag{5.4}$$

and

$$t_{\max} = \max_{\boldsymbol{x}^*\boldsymbol{x}=1} \boldsymbol{x}^*S_2\boldsymbol{x} \leq C_2 \max_{\boldsymbol{x}^*\boldsymbol{x}=1} \boldsymbol{x}^*S_1\boldsymbol{x} = C_2 s_{\max} \tag{5.5}$$

Now, using the fact that $X+Y$ is symmetric, (5.4) and (5.5) give

$$\begin{aligned}
\sigma_{\min}(X+Y) &= \min_{\boldsymbol{x}^*\boldsymbol{x}=1} \boldsymbol{x}^* \left(I - p_s\left[(\alpha_1 S_1 + p_1 hI)^{-1} + (\alpha_2 S_2 + p_2 hI)^{-1}\right]\right)\boldsymbol{x} \\
&\geq 1 - p_s\left[\max_{\boldsymbol{x}^*\boldsymbol{x}=1} \boldsymbol{x}^*(\alpha_1 S_1 + p_1 hI)^{-1}\boldsymbol{x} + \max_{\boldsymbol{x}^*\boldsymbol{x}=1} \boldsymbol{x}^*(\alpha_2 S_2 + p_2 hI)^{-1}\boldsymbol{x}\right] \\
&= 1 - p_s\left[\max_{s\in\sigma(S_1)}\left(\frac{1}{\alpha_1 s + p_1 h}\right) + \max_{t\in\sigma(S_2)}\left(\frac{1}{\alpha_2 t + p_2 h}\right)\right] \\
&= 1 - p_s\left(\frac{1}{\alpha_1 s_{\min} + p_1 h} + \frac{1}{\alpha_2 t_{\min} + p_2 h}\right) \\
&\geq \underbrace{1 - p_s\left(\frac{1}{\alpha_1 s_{\min} + p_1 h} + \frac{1}{C_1\alpha_2 s_{\min} + p_2 h}\right)}_{\mu_1(p_1,p_2)}
\end{aligned}$$

and

$$\begin{aligned}
\sigma_{\max}(X+Y) &= \max_{\boldsymbol{x}^*\boldsymbol{x}=1} \boldsymbol{x}^* \left(I - p_s\left[(\alpha_1 S_1 + p_1 hI)^{-1} + (\alpha_2 S_2 + p_2 hI)^{-1}\right]\right)\boldsymbol{x} \\
&\leq 1 - p_s\left[\min_{\boldsymbol{x}^*\boldsymbol{x}=1} \boldsymbol{x}^*(\alpha_1 S_1 + p_1 hI)^{-1}\boldsymbol{x} + \min_{\boldsymbol{x}^*\boldsymbol{x}=1} \boldsymbol{x}^*(\alpha_2 S_2 + p_2 hI)^{-1}\boldsymbol{x}\right] \\
&= 1 - p_s\left[\min_{s\in\sigma(S_1)}\left(\frac{1}{\alpha_1 s + p_1 h}\right) + \min_{t\in\sigma(S_2)}\left(\frac{1}{\alpha_2 t + p_2 h}\right)\right] \\
&= 1 - p_s\left(\frac{1}{\alpha_1 s_{\max} + p_1 h} + \frac{1}{\alpha_2 t_{\max} + p_2 h}\right) \\
&\leq \underbrace{1 - p_s\left(\frac{1}{\alpha_1 s_{\max} + p_1 h} + \frac{1}{C_2\alpha_2 s_{\max} + p_2 h}\right)}_{\mu_2(p_1,p_2)},
\end{aligned}$$

then we have that $\rho(X+Y) \leq \mathcal{R}(p_1,p_2) = \max\{|\mu_1(p_1,p_2)|\,,\,|\mu_2(p_1,p_2)|\}$.

Again using the fact that $X-Y$ is symmetric, similar calculations as above yield that $\rho(X-Y) \leq \mathcal{I}(p_1,p_2)$. □

As $\mathcal{R}(p_1,p_2)$ and $\mathcal{I}(p_1,p_2)$ are functions of p_1 and p_2, by choosing suitable Robin parameters we hope to make $\mathbf{R}$ "well conditioned" in the sense that GMRES converges quickly. From (4.33) we see that convergence will be quicker if $\mathbf{R}$ is small and far away from the origin. In principle to achieve this we need to minimise both $\mathcal{R}(p_1,p_2)$ and $\mathcal{I}(p_1,p_2)$ however, since $\mathbf{R}$ only approaches the origin along the real axis we choose to focus on minimising $\mathcal{R}(p_1,p_2)$. We are still interested in $\mathcal{I}(p_1,p_2)$, to ensure our choice of parameters doesn't cause $\mathbf{R}$ to be too large in the imaginary direction.

5.2 Optimised Robin parameters

5.2.1 One-sided Robin parameters

Our first choice of Robin parameters is the simple case when we have the same parameters on both sides of the interface, $p_1 = p_2 = q$. Then we have to minimise $\mathcal{R}^{[1]}(q) = \max\{|\mu_1^{[1]}(q)|, |\mu_2^{[1]}(q)|\}$, where

$$\mu_1^{[1]}(q) = 1 - qh\left(\frac{1}{\alpha_1 s_{\min} + qh} + \frac{1}{C_1\alpha_2 s_{\min} + qh}\right)$$

and

$$\mu_2^{[1]}(q) = 1 - qh\left(\frac{1}{\alpha_1 s_{\max} + qh} + \frac{1}{C_2\alpha_2 s_{\max} + qh}\right).$$

We also have that $\mathcal{I}^{[1]}(q) = \max\{|\nu_1^{[1]}(q)|, |\nu_2^{[1]}(q)|\}$, with

$$\nu_1^{[1]}(q) = qh\left(\frac{1}{C_2\alpha_2 s_{\max} + qh} - \frac{1}{\alpha_1 s_{\min} + qh}\right)$$

and

$$\nu_2^{[1]}(q) = qh\left(\frac{1}{C_1\alpha_2 s_{\min} + qh} - \frac{1}{\alpha_1 s_{\max} + qh}\right)$$

This choice of one-sided parameter not only simplifies the analysis but guarantees that both $\mathcal{R}^{[1]}(q) < 1$ and $\mathcal{I}^{[1]}(q) < 1$ for all $q > 0$. Then $W(A_{2LM})$ does not contain the origin and our estimate of (4.33) will hold.

Theorem 5.2.1. (Optimised Robin parameter: one-sided) *The unique minimiser, $q^* > 0$, of $\mathcal{R}^{[1]}(q)$ is given by the solution of*

$$\mu_1^{[1]}(q) = -\mu_2^{[1]}(q). \tag{5.6}$$

Proof. Taking partial derivatives of $\mu_1^{[1]}(q)$ and $\mu_2^{[1]}(q)$ with respect to q we have that

$$\frac{\partial \mu_1^{[1]}}{\partial q} = -\frac{(C_1^2 a_1 a_2^2 s_{\min}^2 + C_1 a_2 q^2 h^2 + 4\, C_1 a_1 a_2 q h s_{\min} + C_1 a_1^2 a_2 s_{\min}^2 + a_1 q^2 h^2) h s_{\min}}{(C_1 a_2 s_{\min} + qh)^2 (qh + a_1 s_{\min})^2} < 0$$

and

$$\frac{\partial \mu_2^{[1]}}{\partial q} = -\frac{(C_2^2 a_1 a_2^2 s_{\max}^2 + C_2 a_2 q^2 h^2 + 4\, C_2 a_1 a_2 q h s_{\max} + C_2 a_1^2 a_2 s_{\max}^2 + a_1 q^2 h^2) h s_{\max}}{(C_2 a_2 s_{\max} + qh)^2 (qh + a_1 s_{\max})^2} < 0$$

for all $q > 0$. Moreover $\mu_j^{[1]}(0) = 1$ and $\lim_{q\to\infty} \mu_j^{[1]}(q) = -1$ for $j = 1, 2$. Then since $\mu_j^{[1]}(q)$ is a monotonically decreasing, continuous function $|\mu_j^{[1]}(q)|$ reaches its minimum when $\mu_j^{[1]}(q) = 0$. Solving $\mu_j^{[1]}(q) = 0$ for q we find that $|\mu_1^{[1]}(q)|$ reaches its minimum at

$$q_1 = \frac{\sqrt{C_1 \alpha_1 \alpha_2}\, s_{\min}}{h},$$

while $|\mu_2^{[1]}(q)|$ reaches its minimum at

$$q_2 = \frac{\sqrt{C_2 \alpha_1 \alpha_2}\, s_{\max}}{h}.$$

It follows that the minimiser q^* of $\mathcal{R}^{[1]}(q)$ must lie in the interval $[q_1, q_2]$. To see this, note that when $q < q_1$ increasing q uniformly decreases both $|\mu_1^{[1]}(q)|$ and $|\mu_2^{[1]}(q)|$. On the other hand if $q > q_2$ decreasing q uniformly decreases both $|\mu_1^{[1]}(q)|$ and $|\mu_2^{[1]}(q)|$, see Figure 5.1.

Now in the interval $[q_1, q_2]$, $|\mu_1^{[1]}(q)|$ is monotonically increasing and $\mu_2^{[1]}(q)$ is monotonically decreasing, so $\mathcal{R}^{[1]}(q)$ must reach its minimum when

$$|\mu_1^{[1]}(q)| = |\mu_2^{[1]}(q)|,$$

i.e. when

$$\mu_1^{[1]}(q) = -\mu_2^{[1]}(q).$$

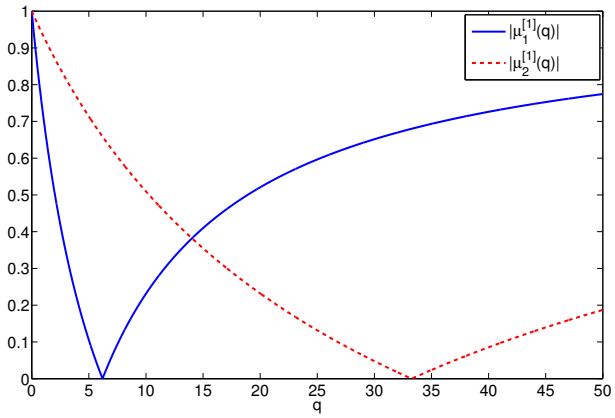

Figure 5.1: upper bound $\mathcal{R}^{[1]}(q)$ for different values of the one-sided Robin parameter q

□

Calculating the minimising Robin parameter in (5.6) involves solving a quartic equation. To simplify the convergence analysis we propose to choose an approximation to q^* by taking the geometric mean of the endpoints of the interval $[q_1, q_2]$. Then we obtain the one-sided Robin parameter:

$$\hat{q} = \frac{\sqrt{\sqrt{C_1 C_2}\alpha_1\alpha_2 s_{\min} s_{\max}}}{h}. \tag{5.7}$$

Let $\mathbf{R}_{\hat{q}}$ denote the rectangle defined by $\mathcal{R}^{[1]}(\hat{q})$ and $\mathcal{I}^{[1]}(\hat{q})$. To approximate the convergence speed of GMRES with this choice of Robin parameter we must construct a conformal map, from the exterior of $\mathbf{R}_{\hat{q}}$ to the interior of the unit disc. We cannot state such a map explicitly but can define its inverse.

Lemma 5.2.2. *Let $\Psi : w \mapsto z$ denote a conformal map from the interior of the unit disc to the exterior of $\mathbf{R}_{\hat{q}}$, with $\Psi(0) = \infty$. Furthermore let $\delta \in (0, 1/2)$ denote the distance, measured along the real axis, from the origin to the left hand boundary of $\mathbf{R}_{\hat{q}}$. Then Ψ is of the form:*

$$\Psi(w, \delta) = \delta + C(\psi(w, \theta) - \psi(1, \theta)), \tag{5.8}$$

where

$$\psi(w,\theta) = \int^{w} \zeta^{-2} \left((1-e^{i\theta}\zeta)(1-e^{-i\theta}\zeta)(1+e^{i\theta}\zeta)(1+e^{-i\theta}\zeta)\right)^{1/2} d\zeta. \tag{5.9}$$

Here $\theta \in (0,\pi/2)$ determines the aspect ratio and $C \in \mathbb{R}^+$ the scaling of $\mathbf{R}_{\hat{q}}$.

Proof. Let $\Xi(w)$ denote a conformal map from the interior of the unit disc to the exterior of an arbitrary rectangle. A *Schwarz-Christoffel mapping* can be used to construct such a map of the form:

$$\Xi(w) = A + C\int^{w} \zeta^{-2} \prod_{k=1}^{4} \left(1 - \frac{\zeta}{w_k}\right)^{1/2} d\zeta,$$

where complex constants A and C correspond to translation and scaling/rotation of the rectangle respectively. The w_k's in the integrand are pre-vertices, where $\Xi(w_k) = z_k$, chosen on the boundary of the unit disc to determine the side lengths of the rectangle. For full details of Schwarz-Christoffel mappings see [9].

Rectangle $\mathbf{R}_{\hat{q}}$ is situated in the right-half plane, centred at the point $(1/2, 0)$ and with sides parallel to the imaginary axis. For a given angle $\theta \in (0,\pi/2)$ consider the choice of pre-vertices $w_1 = e^{-i\theta}$, $w_2 = e^{i\theta}$, $w_3 = -e^{-i\theta}$ and $w_4 = -e^{i\theta}$. Then (5.9) gives a map from the interior of the unit disc to the exterior of the rectangle centred at the origin with sides parallel to the imaginary axis. The choice of θ will determine the aspect ratio of the rectangle, with $\theta = 0$ giving an interval of the real axis, $\theta = \pi/4$ a square and $\theta = \pi/2$ an interval of the imaginary axis.

Now consider the mapping

$$\Psi(w) = A + C\psi(w,\theta). \tag{5.10}$$

Then, for suitably chosen A, C and θ, (5.10) gives a map to the exterior of $\mathbf{R}_{\hat{q}}$, where we need only take A and C to be real and positive.

To this end we can eliminate one of the three unknowns by observing that mapping

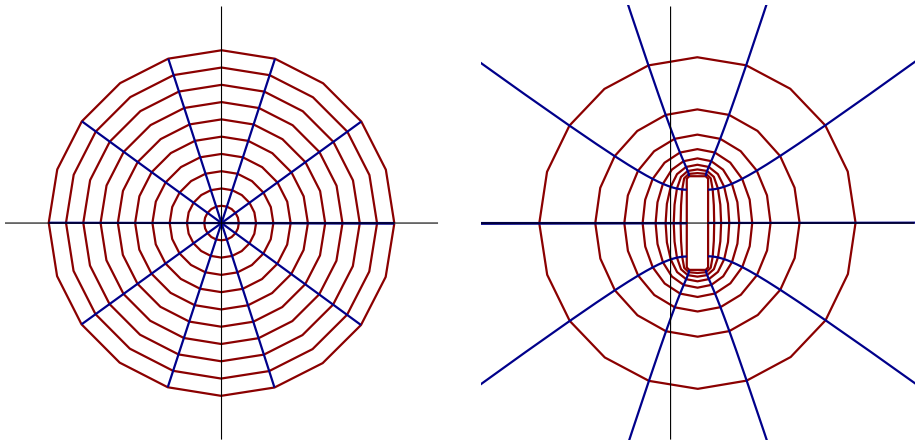

Figure 5.2: conformal mapping from the interior of the unit disc to the exterior of $\mathbf{R}_{\hat{q}}$

(5.9) takes the point $w = 1$ on the boundary of the unit disc to the left hand boundary point, on the real axis, of the rectangle. We know, for $\mathbf{R}_{\hat{q}}$, that the distance of this point from the origin is given by $\delta = \frac{1}{2} - \frac{1}{2}\mathcal{R}^{[1]}(\hat{q})$ and so

$$\delta = A + C\psi(1, \theta).$$

Solving the above for A and substituting into (5.10) we recover the conformal map (5.8) to the exterior of $\mathbf{R}_{\hat{q}}$, as shown in Figure 5.2. □

We are interested in the convergence speed of GMRES when h becomes small, i.e. the finite element mesh size becomes finer, and when the jump in diffusion coefficients α_1 and α_2 becomes large.

Theorem 5.2.3. (Asymptotic performance) *Let the 2LM method system (1.3), with choice of Robin parameter (5.7), be solved using GMRES.*

(i) Assume that α_1 and α_2 are held constant. Let $s_{\max} = C_3/h$ and $h \to 0$. The estimated asymptotic convergence factor of GMRES for small h is

$$\rho = 1 - O(\sqrt{h}). \tag{5.11}$$

(ii) Assume that α_1 is held constant and h is small and held constant. Let $\alpha_2 \to 0$ the estimated asymptotic convergence factor of GMRES for small α_2 is

$$\rho = \beta + O(\sqrt{\alpha_2}), \tag{5.12}$$

where β is a constant, with $\beta < 1$.

Proof. Let $\Phi(z, \delta)$ denote a conformal map from the exterior of $\mathbf{R}_{\hat{q}}$ to the interior of the unit disc. Then (5.8) gives the inverse of $\Phi(z, \delta)$. Taking the linear approximation of $\Psi(w, \delta)$ near $w = 1$, denoted $\overline{\Psi}(w, \delta)$, we obtain

$$\overline{\Psi}(w, \delta) = \delta + \frac{c_1}{c_2} C(w - 1),$$

where

$$c_1 = 8\cos^2\theta\cos 2\theta + 8\sin\theta\cos t\sin 2\theta - \sin 4\theta\sqrt{2 - 2\cos 4\theta} - 7\cos 2\theta - \cos 6\theta$$

and

$$c_2 = \sqrt{2 - 2\cos 2\theta}(1 + \cos 4\theta + \sin 2\theta\sqrt{2 - 2\cos 4\theta}).$$

Solving $\overline{\Psi}(w, \delta) = 0$ for w gives the linear approximation of the mapping from the origin, exterior to $\mathbf{R}_{\hat{q}}$, to the interior of the unit disc:

$$\overline{\Phi}(0, \delta) = 1 - \frac{c_2}{c_1}\frac{\delta}{C}. \tag{5.13}$$

Then from (4.33) we have that the estimated asymptotic convergence factor of GMRES is $\rho = \overline{\Phi}(0, \delta)$, where we have dropped the absolute value since, for suitable C and θ, $\overline{\Phi}(0, \delta)$ is real and positive.

From (5.7) we see that as $\alpha_2 \to 0$, $\hat{q} \to 0$. It follows, observing Fig 5.1, in this case

that $|\mu_2^{[1]}(\hat{q})| > |\mu_1^{[1]}(\hat{q})|$ and $\mu_2^{[1]}(\hat{q}) > 0$. Then we take δ to be of the form

$$\delta = \frac{1}{2} - \frac{1}{2}\mu_2^{[1]}(\hat{q}). \tag{5.14}$$

Substituting (5.14) into (5.13) and taking the series expansion as α_2 goes to zero gives the second result:

$$\begin{aligned}\rho &= \beta \\ &+ \frac{1}{2}\frac{(C_2 s_{\max} - \sqrt{C_1 C_2}s_{\min})\sqrt{2-2\cos 2\theta}(1+\cos 4\theta + \sin 2\theta\sqrt{2-2\cos 4\theta})}{\sqrt{\sqrt{C_1 C_2}\alpha_1 s_{\min} s_{\max}}C(4-3\cos 2\theta - \sin 4\theta\sqrt{2-2\cos 4\theta} - \cos 6\theta))}\sqrt{\alpha_2} \\ &+ O(\alpha_2),\end{aligned}$$

where

$$\begin{aligned}\beta = \frac{1}{2}&\frac{\sqrt{2-2\cos 4\theta}(2C\sin 4\theta + \sin 2\theta\sqrt{2-2\cos 2\theta})}{C(\cos 6\theta + \sin 4\theta\sqrt{2-2\cos 4\theta} + 3\cos 2\theta - 4)} \\ &+ \frac{1}{2}\frac{\sqrt{2-2\cos 2\theta}(1+\cos 4\theta) + C(6\cos 2\theta + 2\cos 6\theta - 8)}{C(\cos 6\theta + \sin 4\theta\sqrt{2-2\cos 4\theta} + 3\cos 2\theta - 4)},\end{aligned}$$

with $\beta < 1$.

Again observing (5.7) we see that as $h \to 0$, $\hat{q} \to \infty$. It follows that $|\mu_1^{[1]}(\hat{q})| > |\mu_2^{[1]}(\hat{q})|$ and $\mu_1^{[1]}(\hat{q}) < 0$. Then we now take δ to be of the form

$$\delta = \frac{1}{2} + \frac{1}{2}\mu_1^{[1]}(\hat{q}). \tag{5.15}$$

It is known, [3], for small h that there exists constant C_3, independent of h, such that $s_{\max} = C_3/h$. Substituting this and (5.15) into (5.13) and taking the series expansion as h

goes to zero gives the first result:

$$\begin{aligned}\rho =&1\\ &-\frac{1}{2}\frac{(\alpha_1+C_1\alpha_2)\sqrt{s_{\min}}\sqrt{2-2\cos 2\theta}(1+\cos 4\theta+\sin 2\theta\sqrt{2-2\cos 4\theta})}{\sqrt{\sqrt{C_1C_2}C_3\alpha_1\alpha_2}C(4-3\cos 2\theta-\sin 4\theta\sqrt{2-2\cos 4\theta}-\cos 6\theta)}\sqrt{h}\\ &+O(h).\end{aligned}$$

□

The estimated asymptotic convergence factor for one-sided parameters behaves like $1-O(\sqrt{h})$ as we refine the mesh, the same behaviour we observed for the OSM iteration with one-sided parameters. However as we increase the jump in diffusion coefficients we observe from (5.12) that the speed of convergence improves, whereas for the OSM iteration with one-sided parameters increasing the jump caused performance to deteriorate. The difference arises in that we are solving the 2LM method system with the GMRES algorithm while the OSM is a fixed point iteration. If we had solved the 2LM system with the Richardson iteration (3.12), we would expect the performance to be the same as the OSM iteration that it its equivalent to. We confirm this in the numerical experiments at the end of the chapter.

5.2.2 Scaled one-sided Robin parameters

Following the choice of scaled one-sided parameters for the OSM in Chapter 2 we now let $p_1=\alpha_2 r$ and $p_2=\alpha_2 r$. The downside to choosing $p_1\neq p_2$ is that for some values of r the field of values of A_{2LM} may contain the origin and so (4.33) will not hold.

We need to minimise $\mathcal{R}^{[1.5]}(r)=\max\{|\mu_1^{[1.5]}(r)|,\,|\mu_2^{[1.5]}(r)|\}$, where

$$\mu_1^{[1.5]}(r)=1-\frac{(\alpha_1+\alpha_2)rh}{2}\left(\frac{1}{\alpha_1 s_{\min}+\alpha_2 rh}+\frac{1}{C_1\alpha_2 s_{\min}+\alpha_1 rh}\right)$$

and

$$\mu_2^{[1.5]}(r) = 1 - \frac{(\alpha_1+\alpha_2)rh}{2}\left(\frac{1}{\alpha_1 s_{\max}+\alpha_2 rh} + \frac{1}{C_2\alpha_2 s_{\max}+\alpha_1 rh}\right).$$

Theorem 5.2.4. (Optimised Robin parameter: scaled one-sided) *The unique minimiser, $r^* > 0$, of $\mathcal{R}^{[1.5]}(r)$ is given by the solution of*

$$\mu_1^{[1.5]}(r) = -\mu_2^{[1.5]}(r). \tag{5.16}$$

Proof. We proceed in a similar fashion as we did for Theorem 5.2.1. Taking the partial derivatives of $\mu_1^{[1.5]}(r)$ and $\mu_2^{[1.5]}(r)$ with respect to r we see that

$$\begin{aligned}\frac{\partial \mu_1^{[1.5]}}{\partial r} &= -\frac{\left(C_1 a_2^3 q^2h^2 + a_1^3q^2h^2 + 2C_1a_1^2a_2qhs_{\min} + 2C_1a_1a_2^2qhs_{\min} + C_1^2a_1a_2^2s_{\min}^2 + C_1a_1^2a_2s_{\min}^2\right)(a_1+a_2)hs_{\min}}{2(a_1qh + C_1a_2s_{\min})^2(a_2qh+a_1s_{\min})^2}\\ &< 0\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \mu_2^{[1.5]}}{\partial r} &= -\frac{\left(C_2 a_2^3 q^2h^2 + a_1^3q^2h^2 + 2C_2a_1^2a_2qhs_{\max} + 2C_2a_1a_2^2qhs_{\max} + C_2^2a_1a_2^2s_{\max}^2 + C_2a_1^2a_2s_{\max}^2\right)(a_1+a_2)hs_{\max}}{2(a_1qh + C_2a_2s_{\max})^2(a_2qh+a_1s_{\max})^2}\\ &< 0\end{aligned}$$

for all $r > 0$. Moreover $\mu_j^{[1.5]}(0) = 1$ and $\lim_{r\to\infty}\mu_j^{[1.5]}(r) = -\frac{1}{2}\frac{\alpha_1^2+\alpha_2^2}{\alpha_1\alpha_2}$, for $j = 1, 2$. Let

$$D_j = \alpha_1^4 + (6C_j-2)\alpha_1^3\alpha_2 + (C_j^2+4C_j+1)\alpha_1^2\alpha_2^2 + (6C_j-2C_j^2)\alpha_1\alpha_2^2 + C_j\alpha_2^4,$$

then $|\mu_1^{[1.5]}(r)|$ reaches its minimum at

$$r_1 = \frac{\left(\alpha_1^2 - (C_1+1)\alpha_1\alpha_2 + C_1\alpha_2^2 + \sqrt{D_1}\right)s_{\min}}{2(\alpha_1^2+\alpha_2^2)h}$$

and $|\mu_2^{[1.5]}(r)|$ reaches its minimum at

$$r_2 = \frac{\left(\alpha_1^2 - (C_2+1)\alpha_1\alpha_2 + C_2\alpha_2^2 + \sqrt{D_2}\right) s_{\max}}{2(\alpha_1^2+\alpha_2^2)h}.$$

It follows that since $\mu_1^{[1.5]}(r)$ and $\mu_2^{[1.5]}(r)$ are monotonically decreasing functions the minimiser r^* must lie in the interval $[r_1,\, r_2]$.

Then we have that in the interval $[r_1, r_2]$, $|\mu_1^{[1.5]}(r)|$ is monotonically increasing and $|\mu_2^{[1.5]}(r)|$ is monotonically decreasing. So the unique minimiser r^* is obtained when

$$|\mu_1^{[1.5]}(r)| = |\mu_2^{[1.5]}(r)|,$$

i.e. when

$$\mu_1^{[1.5]}(r) = -\mu_2^{[1.5]}(r).$$

□

The minimising scaled one-sided Robin parameter r^* results in $W(A_{2LM})$ containing the origin, so we cannot derive a convergence estimate of the form (4.33). However we will see in the numerical experiments in the next section that this choice of parameters leads to faster convergence as compared to the non-scaled one-sided case.

As with the OSM it is also possible to consider two-sided Robin parameters, where p_1 and p_2 are completely independent of each other. In the derivation of the 2LM method and the approximation of its field of values by a rectangle we have allowed p_1 and p_2 to be independent and, apart from being non-negative, completely arbitrary. In the special cases of one-sided and scaled one-sided parameters we have seen above that the formula for the rectangle encasing the field of values simplified just enough that we could analytically find the optimised parameters.

As was shown in Chapter 2 these parameters can be derived for the OSM in the special

case that the subdomains are rectangular and symmetric about the interface. In the case of general domains and subdomains, two-sided parameters are difficult to derive. When we allow p_1 and p_2 to vary independently the analogue of Figure 5.1 would be a surface and we were not able to find analytically the corresponding optimal parameters. If the two-sided parameters could be found as in the case of the OSM one would expect them to perform better than either one-sided or scaled one-sided parameters.

5.3 Numerical experiments

We consider model problem (1.2) on the L-shaped domain $\Omega \subset \mathbb{R}^2$, which is partitioned into two general non-overlapping subdomains as shown in Figure 5.3. The shape of the subdomains and the interface Γ that separates them is chosen to emphasise that our results for the 2LM method hold for general subdomains. This is in contrast to the results derived for the OSM in Chapter 2, where a model problem with a straight interface was used.

The diffusion coefficient is given by

$$a(x,y) = \begin{cases} \alpha_1(x,y)(1+\frac{1}{2}\sin(3\pi x)\cos(3\pi y)) & \text{in} \quad \Omega_1 \\ \\ \alpha_2(x,y)(1+\frac{1}{2}\sin(3\pi x)\cos(3\pi y)) & \text{in} \quad \Omega_2, \end{cases}$$

where

$$\alpha_1(x,y) = 1 + \frac{1}{2}\sin(3\pi xy)$$

and

$$\alpha_2(x,y) = \omega\left(1 + \frac{1}{2}\cos\left(\frac{3\pi xy}{\omega^{1/4}}\right)\right),$$

here ω is a constant such that as ω decreases the jump in coefficients between subdomains

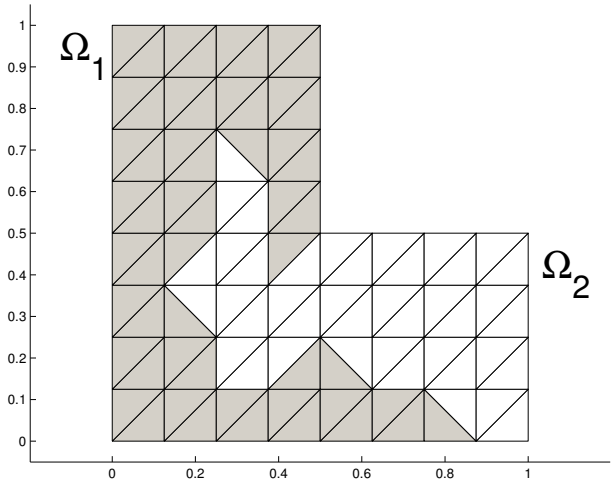

Figure 5.3: non-overlapping decomposition of an L-shaped domain into two general subdomains

increases. The forcing term is given by

$$f(x,y) = 1 + \frac{1}{2}\alpha_1(x,y)\alpha_2(x,y)\sin(3\pi x)\sin(3\pi y).$$

Note that in the example above the diffusion coefficients α_1 and α_2 vary within their respective subdomains and along the interface, while our analysis assumed these coefficients were constant. However, we shall see that with a suitable choice of Robin parameters numerical results similar to our theoretical results hold.

We perform a uniform triangulation of Ω, with mesh parameter h, and discretise the PDE using piecewise linear, triangular finite elements. We solve system (1.3) using the built in GMRES solver in MATLAB for different choices of h and ω. We use GMRES without restarts, a zero vector initial guess and are interested in the number of iterations required to reach a tolerance of 10^{-12} in the relative residual. We also calculate the 2-norm condition number of A_{2LM} using the MATLAB command `cond`.

First we consider the case of non-scaled one-sided parameters. Note that our choice of

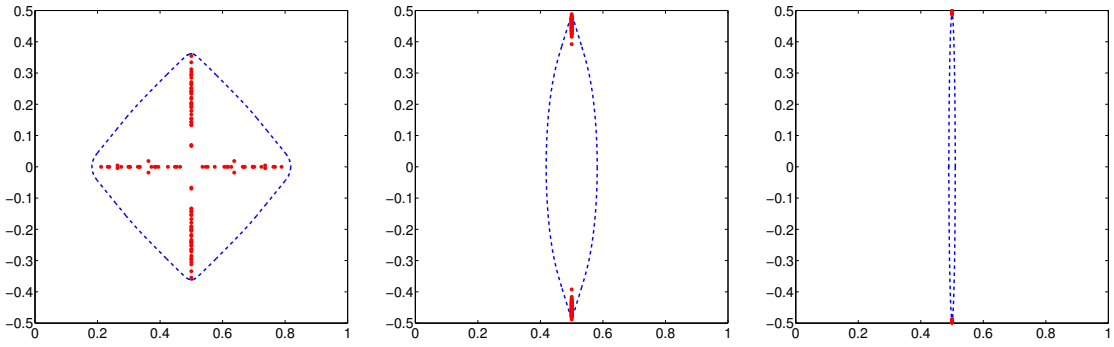

Figure 5.4: $W(A_{2LM})$ (dashed line) and $\sigma(A_{2LM})$ (dots) with choice of one-sided Robin parameter (5.17), $h = 1/32$, $\omega = 10^{-1}$ on left, $\omega = 10^{-3}$ in middle and $\omega = 10^{-5}$ on right

one-sided parameters given in (5.7) involves constants C_1 and C_2 that, in general, we do not know. Then we choose to set $C_1 = C_2 = 1$, which corresponds to the case of symmetric subdomains about the interface. We do not have symmetric subdomains for our example, but as we will see even with this simplification we can still achieve fast convergence. Moreover our choice of one-sided parameters given in (5.7) assumes the coefficients α_1 and α_2 are constant whereas in our example they vary within their subdomains and along the interface. Then if $\boldsymbol{\alpha}_i$ denotes the vector of values of $\alpha_i(x, y)$ corresponding to our discretisation in Ω_i we propose to take the arithmetic mean of these vectors. The one-sided Robin parameter used is

$$\tilde{q} = \frac{\sqrt{\beta_1 \beta_2 s_{\min} s_{\max}}}{h}, \tag{5.17}$$

where β_1 and β_2 denote the arithmetic mean of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ respectively.

Eigenvalues $s_{\min}$ and $s_{\max}$ are calculated in MATLAB using the `eigs` command. For larger problems where this approach would be impractical, due to S_1 and S_2 being dense, the estimates $s_{\min} = b_1$ and $s_{\max} = b_2/h$, for constants b_1 and b_2, (see [3]) can be used. The results for the one-sided Robin parameter are shown in Table 5.1 for different values of h and ω.

We see that even though our example has non-constant coefficients and we have sim-

	$h = 1/16$	$h = 1/32$	$h = 1/64$	$h = 1/128$
$\omega = 10^{-1}$	**27** (2.9228)	**33** (4.1869)	**38** (6.7926)	**46** (10.2553)
$\omega = 10^{-2}$	**22** (1.6039)	**24** (1.9787)	**28** (2.5620)	**32** (3.6564)
$\omega = 10^{-3}$	**16** (1.1995)	**18** (1.3142)	**20** (1.5109)	**24** (1.7949)
$\omega = 10^{-4}$	**12** (1.0631)	**14** (1.0998)	**16** (1.1507)	**16** (1.2235)
$\omega = 10^{-5}$	**10** (1.0218)	**12** (1.0309)	**12** (1.0470)	**12** (1.0691)

Table 5.1: number of iterations of GMRES (in bold) and the condition number of A_{2LM} matrix (in brackets) using one-sided Robin parameter (5.17)

plified the optimised Robin parameter the numerical results confirm the theoretical results from Theorem 5.2.3. Decreasing the mesh size h requires more iterations of GMRES while increasing the jump in coefficients requires less. Though in general the condition number of a non-symmetric matrix is not useful in determining the speed of convergence of GMRES here we see favourable behaviour of the condition number of A_{2LM} with one-sided parameters.

In Figure 5.4 we plot the spectrum and field of values of A_{2LM} for different values of ω with one-sided parameters. We see that as the jump in coefficients becomes larger the field of values, which is a close approximation of the convex hull of the eigenvalues, moves further away from the origin while the eigenvalues become more clustered together in two points. This clustering causes the field of values to become "skinnier" with its boundary away from the origin, then GMRES can be expected to perform reasonably well as predicted by our theoretical results.

For the case of scaled one-sided parameters, as in the case of one-sided parameters, the minimising parameter obtained from solving (5.16) involves the solution of a quartic equation. So we choose to approximate the parameter by taking the geometric mean of

	$h = 1/16$	$h = 1/32$	$h = 1/64$	$h = 1/128$
$\omega = 10^{-1}$	**19** (10.9906)	**22** (22.5532)	**26** (43.7648)	**29** (77.8994)
$\omega = 10^{-2}$	**12** (19.3961)	**12** (45.7221)	**14** (103.6398)	**15** (217.4882)
$\omega = 10^{-3}$	**9** (21.0156)	**10** (50.6331)	**10** (118.0515)	**10** (256.2361)
$\omega = 10^{-4}$	**6** (21.2003)	**6** (51.1948)	**8** (119.7098)	**8** (260.8631)
$\omega = 10^{-5}$	**6** (21.2191)	**6** (51.2524)	**6** (119.8828)	**6** (261.3467)

Table 5.2: number of iterations of GMRES (in bold) and the condition number of A_{2LM} matrix (in brackets) using scaled one-sided Robin parameter (5.18)

the endpoints of the interval $[r_1, r_2]$, as defined in Theorem 5.2.4, in which r^* lies. Again we set $C_1 = C_2 = 1$ and let β_1 and β_2 be the arithmetic mean of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ respectively. Then the scaled one-sided Robin parameter we use is given by

$$\tilde{r} = \frac{\sqrt{s_{\min} s_{\max}}}{h} \tag{5.18}$$

and the Robin parameters used for system (1.3) are $p_1 = \beta_2 \tilde{r}$ and $p_2 = \beta_1 \tilde{r}$. The results for this choice for different values of h and ω are shown in Table 5.2 while the spectrum and field of values of A_{2LM} for different values of ω are shown in Figure 5.5.

For scaled one-sided parameters we observe similar behaviour as we do for the non-scaled parameters. The number of iterations increases as we decrease h and decreases as we decrease ω. However, the number of iterations needed to reach the same tolerance is significantly less. For small h and ω we require about half the number of iterations as are required for the non-scaled case.

Figure 5.5 goes some way to explain this. In contrast to the one-sided case, as ω decreases the size of $W(A_{2LM})$ increases and the field of values becomes much larger than

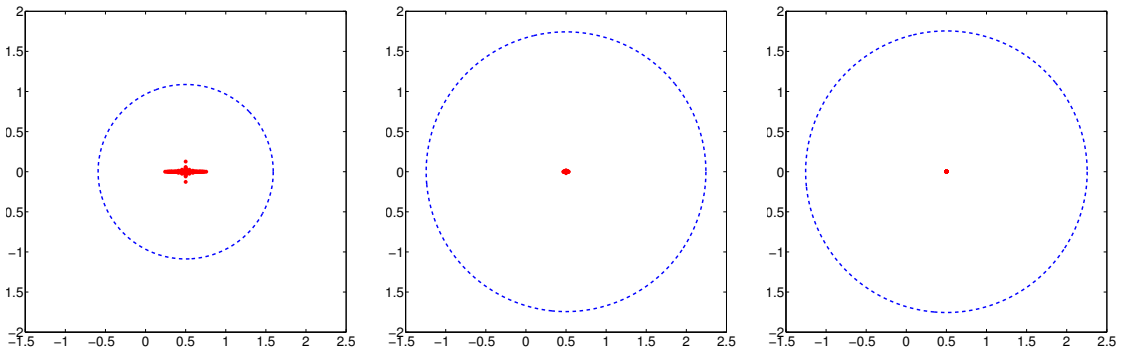

Figure 5.5: $W(A_{2LM})$ (dashed line) and $\sigma(A_{2LM})$ (dots) with choice of scaled one-sided Robin parameter (5.18), $h = 1/32$, $\omega = 10^{-1}$ on left, $\omega = 10^{-3}$ in middle and $\omega = 10^{-5}$ on right

the convex hull of the eigenvalues. This is due to the high non-normality of matrix A_{2LM} when we have scaled one-sided parameters. The field of values is a good approximation of the spectrum of a matrix when said matrix is not too highly non-normal, [17], as is the case when we have non-scaled one-sided parameters. Despite this we see that as the jump increases the eigenvalues cluster together near a single point on the real line. In this regime we expect GMRES to perform much better than the pessimistic estimate the field of values would indicate, indeed as the field of values contains the origin the field of values does not predict the convergence. Whereas in the non-scaled case the clustering happens around two points here we have one cluster. GMRES can more easily deal with systems whose eigenvalues are clustered together away from the origin and are on the real line. Despite the field of values being a poor estimate in the scaled one-sided case we see from Table 5.2 that the approach of minimising $W(A_{2LM})$ for scaled one-sided parameters results in much faster convergence of GMRES.

We see from the theoretical and numerical results that increasing the jump in coefficients for one-sided parameters used with the 2LM method reduces the iteration count of GMRES. In our numerical results for the OSM iteration with one-sided parameters the iteration count increased as we increased the jump. The reason is that the OSM is a fixed point

iteration while we are solving the 2LM system with GMRES. Recasting the OSM as a linear system and solving using GMRES we should observe similar performance as we did for solving the 2LM system.

From the discrete OSM iteration (2.27) we have

$$\begin{aligned}\begin{bmatrix}\boldsymbol{u}_1^k\\ \boldsymbol{u}_2^k\end{bmatrix} &= \begin{bmatrix}(\alpha_1 A_{N_1}+p_1B_1)^{-1}\left(R_1(\boldsymbol{f}-\alpha_1\tilde{A}R_2^T\boldsymbol{u}_2^{k-1})+(\alpha_2A_{N_1}+p_1B_1)R_1R_2^T\boldsymbol{u}_2^{k-1}\right)\\ (\alpha_2 A_{N_2}+p_2B_2)^{-1}\left(R_2(\boldsymbol{f}-\alpha_2\tilde{A}R_1^T\boldsymbol{u}_1^{k-1})+(\alpha_1A_{N_2}+p_2B_2)R_2R_1^T\boldsymbol{u}_1^{k-1}\right)\end{bmatrix}\\ &= M_1M_2\begin{bmatrix}\boldsymbol{u}_1^{k-1}\\ \boldsymbol{u}_2^{k-1}\end{bmatrix}+M_1\begin{bmatrix}R_1\boldsymbol{f}\\ R_2\boldsymbol{f}\end{bmatrix},\end{aligned}$$

where

$$M_1=\begin{bmatrix}(\alpha_1A_{N_1}+p_1B_1)^{-1} & 0\\ 0 & (\alpha_2A_{N_2}+p_2B_2)^{-1}\end{bmatrix}$$

and

$$M_2=\begin{bmatrix}0 & (\alpha_2A_{N_1}+p_1B_1)R_1R_2^T-\alpha_2R_1\tilde{A}R_2^T\\ (\alpha_1A_{N_2}+p_2B_2)R_2R_1^T-\alpha_1R_2\tilde{A}R_1^T & 0\end{bmatrix}.$$

Let

$$\boldsymbol{U}^k=\begin{bmatrix}\boldsymbol{u}_1^k\\ \boldsymbol{u}_2^k\end{bmatrix}\quad\text{and}\quad\boldsymbol{F}=\begin{bmatrix}R_1\boldsymbol{f}\\ R_2\boldsymbol{f}\end{bmatrix}$$

then we have the Richardson iteration

$$\boldsymbol{U}^k=\boldsymbol{U}^{k-1}+(M_1\boldsymbol{F}-(I-M_1M_2)\boldsymbol{U}^{k-1}),$$

	$h = 1/16$		$h = 1/32$		$h = 1/64$		$h = 1/128$	
Parameter:	1	1.5	1	1.5	1	1.5	1	1.5
$\omega = 10^{-1}$	26	19	29	21	34	23	39	26
$\omega = 10^{-2}$	21	11	21	12	24	13	27	13
$\omega = 10^{-3}$	15	8	15	8	17	9	17	9
$\omega = 10^{-4}$	11	7	11	6	11	6	11	7
$\omega = 10^{-5}$	9	5	9	5	9	5	9	5

Table 5.3: number of iterations for solving the augmented OSM system (5.19) with GMRES, for one-sided (denoted 1) and scaled one-sided (denoted 1.5) parameters

for the linear system

$$(I - M_1 M_2)\boldsymbol{U} = M_1 \boldsymbol{F}, \tag{5.19}$$

which we can solve using the GMRES method. Table 5.3 shows the iteration count for the GMRES method when used to solve the OSM linear system (5.19). The results are similar to those when the 2LM system is solved with the GMRES method. Refining the mesh increases the iteration count for both choices of parameter, while increasing the jump in coefficients improves the performance of both choices with the scaled parameter performing better than the non-scaled parameter.

We have shown in Lemma 3.2.1 that the 2LM system, when solved with a Richardson iteration, is equivalent to the OSM iteration. Then we should expect the same behaviour numerically as was shown in the experiments in Chapter 2. Table 5.4 shows the iteration count for both one-sided and scaled one-sided parameters when solving the 2LM system with a Richardson iteration. As The Richardson iteration is a fixed point iteration the

	$h = 1/16$		$h = 1/32$		$h = 1/64$		$h = 1/128$	
Parameter:	1	1.5	1	1.5	1	1.5	1	1.5
$\omega = 10^{-1}$	43	15	39	20	36	25	42	33
$\omega = 10^{-2}$	126	9	115	11	107	13	105	13
$\omega = 10^{-3}$	376	7	347	7	331	7	328	9
$\omega = 10^{-4}$	>1000	5	>1000	5	>1000	7	>1000	7
$\omega = 10^{-5}$	>1000	5	>1000	5	>1000	5	>1000	5

Table 5.4: number of iterations for solving the augmented 2LM system with a Richardson iteration, for one-sided (denoted 1) and scaled one-sided (denoted 1.5) parameters

stopping criterion is $\|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k-1}\|_2 < 10^{-8}$. We observe the same behaviour as we did for the OSM iteration for both parameters refining the mesh increases the iteration count. Increasing the jump in coefficients causes the performance of the one-sided parameters to deteriorate and the performance of the scaled one-sided parameters to improve.

Chapter 6

The case of many subdomains and cross points

6.1 Formulation

We return to the formulation of the 2LM method and now consider the case of problems with many subdomains. Recall that a *cross point* is a point where three or more subdomains coincide. The results we have seen thus far apply to problems involving two non-overlapping subdomains or those that have multiple non-overlapping subdomains arranged in such a way that no cross points are present, i.e. in strips, see Figure 6.1 on the left for an example. In general applications we are likely to encounter heterogeneous problems that give rise to a natural decomposition of the global domain into many non-overlapping subdomains with cross points, like that on the right of Figure 6.1.

To see how cross points can give rise to difficulties in the analysis of the OSM and 2LM method consider that we want to solve a strong form PDE $-\Delta u = f$ on the two subdomain decomposition shown on the left of Figure 6.2. This decomposition does not have a cross point but has a corner marked by the red point. When applying the OSM or 2LM method

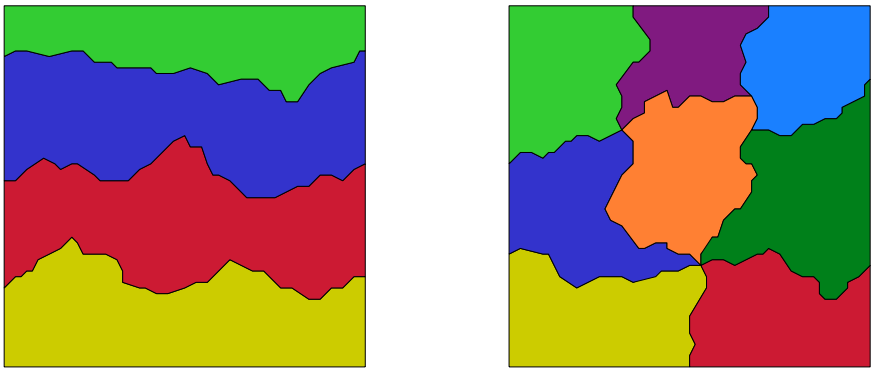

Figure 6.1: multiple non-overlapping subdomains without cross points (left) and with cross points (right)

we need to find the outward pointing normal derivative along the interface between the subdomains. If h is the grid spacing then the normal derivative for the top edge of Ω_1, those points marked in green. is given by

$$\frac{u_1(x+h,y)-u_1(x-h,y)}{2h}.$$

while the normal derivative for the right side edge of Ω_1 is given by

$$\frac{u_1(x,y+h)-u_1(x,y-h)}{2h}.$$

On the corner, marked by the red point, no uniquely defined normal derivative exists, so some "arbitrary" decision must be made.

We can help ourselves by considering the weak form of the PDE, with a test function

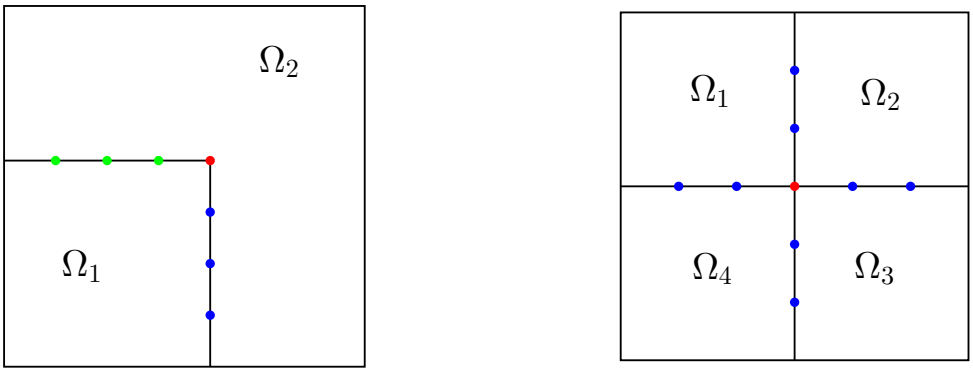

Figure 6.2: unit square decomposed into two non-overlapping subdomains without cross points (left) and into four non-overlapping subdomains with one cross point marked in red (right)

v:

$$\int_{\Omega_1} \nabla u_1 \cdot \nabla v = \int_{\Omega_1} fv + \int_{\Gamma} \lambda_1 v,$$

where λ_1 is the Neumann data for u_1 which we can solve for, or impose, if we know u_1 and f.

On the other hand say we want to solve the same problem but now on the decomposition into four subdomains as shown on the right of Figure 6.2. Now we have a cross point, marked by the red point. Say we wish to know the outward pointing normal derivative of the cross point with respect to subdomain Ω_1. How do we read the Neumann data from subdomains Ω_2, Ω_2 and Ω_3? There is no uniquely defined normal and the weak form will give us three Neumann data corresponding to the other subdomains that could be completely different.

Now some tricky decision must be made at the cross point. Loisel in [45] showed how to do this for the 2LM method for homogeneous problems and we will follow this approach

here for heterogeneous problems with cross points.

To derive the 2LM method for multiple subdomains with cross points let us again consider the heterogeneous problem (1.2) and now assume the domain is decomposed into m non-overlapping subdomains, $\Omega_1, \ldots, \Omega_m$, that may meet each other at cross points. Each of the subdomains has a corresponding diffusion coefficient, $\alpha_1, \ldots, \alpha_m$ that can create discontinuities across the interface $\Gamma = \cup_{i=1}^{m} \partial\Omega_i \setminus \partial\Omega$. After a suitable discretisation we have the linear system

$$A\boldsymbol{u} = \boldsymbol{f} \tag{6.1}$$

The local discrete Robin problem for subdomain i in block form is given by

$$\begin{bmatrix} \alpha_i A_{II_i} & \alpha_i A_{I\Gamma_i} \\ \alpha_i A_{\Gamma I_i} & \alpha_i A_{\Gamma\Gamma_i} + B_i \end{bmatrix} \overbrace{\begin{bmatrix} \boldsymbol{u}_{I_i} \\ \boldsymbol{u}_{\Gamma_i} \end{bmatrix}}^{\boldsymbol{u}_i} = \begin{bmatrix} \boldsymbol{f}_{I_i} \\ \boldsymbol{f}_{\Gamma_i} \end{bmatrix} + \begin{bmatrix} 0 \\ \boldsymbol{\lambda}_i \end{bmatrix}, \tag{6.2}$$

here $B_i = p_i h I$ is the lumped mass matrix and p_i the Robin parameter along the interface Γ_i, the part of the interface Γ that belongs to subdomain i. Eliminating the interior nodes, $\boldsymbol{u}_{I_i}$ from (6.2), we have for the unknowns on the interface:

$$(\alpha_i S_i + B_i)\boldsymbol{u}_{\Gamma_i} = \boldsymbol{g}_i + \boldsymbol{\lambda}_i, \tag{6.3}$$

where $S_i = A_{\Gamma\Gamma_i} - A_{\Gamma I_i} A_{II_i}^{-1} A_{I\Gamma_i}$ and $\boldsymbol{g}_i = \boldsymbol{f}_{\Gamma_i} - A_{\Gamma I_i} A_{II_i}^{-1} \boldsymbol{f}_{I_i}$ are the Schur complement and the accumulated right hand side, respectively.

Now defining the block matrices

$$S = diag\{\alpha_1 S_1, \ldots, \alpha_m S_m\} \qquad \text{and} \qquad B = diag\{B_1, \ldots, B_m\},$$

(6.3) is equivalent to the system

$$(S + B)\boldsymbol{u}_G = \boldsymbol{g} + \boldsymbol{\lambda}. \tag{6.4}$$

Here $\boldsymbol{u}_G = [\boldsymbol{u}_{\Gamma_i}^T, \ldots, \boldsymbol{u}_{\Gamma_m}^T]^T$, $\boldsymbol{g} = [\boldsymbol{g}_i^T, \ldots, \boldsymbol{g}_m^T]^T$ and $\boldsymbol{\lambda} = [\boldsymbol{\lambda}_1^T, \ldots, \boldsymbol{\lambda}_m^T]^T$. For a solution to (6.1) to hold, the many sided trace vector $\boldsymbol{u}_G$ must correspond to a continuous function across Γ. Let K be the orthogonal projection matrix whose range is the space of continuous many sided traces. Then we seek a solution such that

$$K\boldsymbol{u}_G = \boldsymbol{u}_G, \tag{6.5}$$

is satisfied.

Recall that for each subdomain, Ω_i for $i = 1, \ldots, m$, there is an associated matrix R_i that restricts an arbitrary n-dimensional vector to one with entries corresponding to the degrees of freedom in $\Omega_i \cup \Gamma_i$. Partitioning these matrices into blocks corresponding first to degrees of freedom in Ω_i and then those on Γ_i we have

$$R_i = \begin{bmatrix} R_{I_i} \\ R_{\Gamma_i} \end{bmatrix}.$$

Now let

$$R_G = \begin{bmatrix} R_{\Gamma_1} \\ \vdots \\ R_{\Gamma_m} \end{bmatrix}.$$

Then the averaging matrix K is of the form:

$$K = W R_G R_G^T$$

where

$$W = [\text{diag}(R_G R_G^T \mathbf{1})]^{-1},$$

here $\mathbf{1}$ is a vector of all ones.

The 2LM method is again to solve a system of the form:

$$A_{2LM}\boldsymbol{\lambda} = \boldsymbol{c}, \tag{6.6}$$

where now

$$A_{2LM} = (I - BKB^{-1} - K)\overbrace{B(S+B)^{-1}}^{Q} + K \tag{6.7}$$

and

$$\boldsymbol{c} = -(I - BKB^{-1} - K)B(S+B)^{-1}\boldsymbol{g}.$$

When we have just two subdomains $S = \text{diag}\{\alpha_1 S_1,\, \alpha_2 S_2\}$, $B = \text{diag}\{p_1 hI,\, p_2 hI\}$ and $K = \frac{1}{2}\left[\begin{smallmatrix} I & I \\ I & I \end{smallmatrix}\right]$. With a straightforward block matrix calculation we can derive from (6.7) the form of the 2LM system given in Corollary 5.1.2.

Once we have solved the 2LM system we have solutions $\boldsymbol{u_i}$ to the local Robin problems such that

$$\boldsymbol{u}_i = R_i \boldsymbol{u} \quad \text{for} \quad i = 1, \ldots, m. \tag{6.8}$$

Again we can write the global stiffness matrix and load vector in the form:

$$A = \sum_{i=1}^{m} \alpha_i R_i^T A_{N_i} R_i \quad \text{and} \quad \boldsymbol{f} = \sum_{i=1}^{m} R_i^T \boldsymbol{f}_i \tag{6.9}$$

To see how we have reached the form of the 2LM system shown in (6.6) - (6.7), we first have the following result.

Lemma 6.1.1. *Assume that A is non-singular. Let $R_\Gamma = [0 \quad I] \in \mathbb{R}^{(n-n_I)\times n}$ be the matrix*

that restricts solution vector $\boldsymbol{u}$ *to its block component* $\boldsymbol{u}_\Gamma$ *on the interface. There exists a solution* $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ *and* $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m$ *to (6.5), (6.2) and*

$$\sum_{i=1}^{m} R_\Gamma R_{\Gamma_i}^T \boldsymbol{\lambda}_i = \sum_{i=1}^{m} R_\Gamma R_{\Gamma_i}^T B_k \boldsymbol{u}_{\Gamma_i}. \tag{6.10}$$

Furthermore the solution $\boldsymbol{u}$ *to (6.8) solves (6.1).*

Proof. Assume we have a solution $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ and $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m$ that satisfies (6.5), (6.2) and (6.10). From (6.5) the local solution $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ satisfies the continuity condition so there exists a $\boldsymbol{u}$ such that (6.8) holds. Then we seek to show that this $\boldsymbol{u}$ solves (6.1).

We have from (6.9) and (6.5) that

$$\begin{aligned} A\boldsymbol{u} &= \sum_{i=1}^{m} R_i^T \begin{bmatrix} \alpha_i A_{II_i} & \alpha_i A_{I\Gamma_i} \\ \alpha_i A_{\Gamma I_i} & \alpha_i A_{\Gamma\Gamma_i} \end{bmatrix} R_i \boldsymbol{u} \\ &= \begin{bmatrix} A_{II}\boldsymbol{u}_I + A_{I\Gamma}\boldsymbol{u}_\Gamma \\ \sum_{i=1}^{m} R_\Gamma R_{\Gamma_i}^T (\alpha_i A_{\Gamma I_i}\boldsymbol{u}_{I_i} + \alpha_i A_{\Gamma\Gamma_i}\boldsymbol{u}_{\Gamma_i}) \end{bmatrix}. \end{aligned}$$

Now from (6.2) and (6.10) we obtain

$$\begin{aligned} A\boldsymbol{u} &= \begin{bmatrix} \boldsymbol{f}_I \\ \sum_{i=1}^{m} R_\Gamma R_{\Gamma_i}^T (\boldsymbol{f}_{\Gamma_i} + \boldsymbol{\lambda_i} - B_i) \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{f}_I \\ \sum_{i=1}^{m} R_\Gamma R_{\Gamma_i}^T \boldsymbol{f}_{\Gamma_i} \end{bmatrix} \\ &= \boldsymbol{f}, \end{aligned}$$

as required.

To see that the solution $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ and $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m$ is unique, assume we have a different solution $\boldsymbol{u}_1^*, \ldots, \boldsymbol{u}_m^*$ and $\boldsymbol{\lambda}_1^*, \ldots, \boldsymbol{\lambda}_m^*$ to (6.5), (6.2) and (6.10). If $\boldsymbol{u}_i = \boldsymbol{u}_i^*$ for all $i = 1, \ldots, m$, then from (6.2) we must have that $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}_i^*$ for $i = 1, \ldots, m$.

Then assuming $\boldsymbol{u}_i \neq \boldsymbol{u}_i^*$ for some i there is a $\boldsymbol{u}^*$ that satisfies (6.8) and hence $A\boldsymbol{u}^* = \boldsymbol{f}$. Since A is invertible we have that $\boldsymbol{u}^* = A^{-1}\boldsymbol{f} = \boldsymbol{u}$. It follows that $\boldsymbol{u}_i^* = R_i\boldsymbol{u}^* = R_i\boldsymbol{u} = \boldsymbol{u}_i$ which contradicts $\boldsymbol{u}_i^* \neq \boldsymbol{u}_i$. Hence the solution to (6.5), (6.2) and (6.10) is unique. □

From the above lemma we have the following way to recover the solution $\boldsymbol{u}$ to (6.1). Using the formula for $\boldsymbol{u}_G$ provided by (6.4) and the formula for $\boldsymbol{u}_{\Gamma_i}$ provided by (6.3) (on eliminating the interior nodes of (6.2)), systems (6.5) and (6.10) become

$$(I - K)(S + B^{-1})\boldsymbol{\lambda} = (K - I)(S + B)^{-1}\boldsymbol{g} \tag{6.11}$$

and

$$\sum_{i=1}^{m} R_\Gamma R_{\Gamma_i}^T (I - B_i(S_i + B_i)^{-1})\boldsymbol{\lambda}_i = \sum_{i=1}^{m} R_\Gamma R_{\Gamma_i}^T B_i(S_i + B_i)^{-1}\boldsymbol{g}_i, \tag{6.12}$$

respectively.

Solving the system defined above for $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m$ we can recover the global solution $\boldsymbol{u}$ by using (6.2) and (6.8). However while (6.11) is a square system (6.12) is rectangular, in practice it is easier to solve a square non-singular system. We can achieve this by choosing suitable matrices C_1 and C_2 such that the system given by C_1(6.11) + C_2(6.12) is square. The choices

$$C_1 = B \quad \text{and} \quad C_2 = W \begin{bmatrix} R_{\Gamma_1} R_\Gamma^T \\ \vdots \\ R_{\Gamma_m} R_\Gamma^T \end{bmatrix} \tag{6.13}$$

give us the 2LM system defined by (6.7) and (6.6).

To see that matrices C_1 and C_2 are the correct choices and that we can recover the solution to (6.1) from that of (6.6) we have the following result.

Theorem 6.1.2. *Assume A and $S + B$ are non-singular and B^{-1} is positive definite. System (6.6) is equivalent to (6.1).*

Proof. It is enough to show that the rows of the left hand side of (6.11) and (6.12) lie in the

linear span of the rows of (6.7). We start by recovering the rows of (6.12). Left-multiplying A_{2LM} by KB^{-1} and using the fact that K is an orthogonal projection, i.e. $K^2 = K$, we have

$$KB^{-1}A_{2LM} = KB^{-1}K(I - B(S+B)^{-1}). \tag{6.14}$$

By showing that the range of $KB^{-1}K$ is the same as the range of K we can recover (6.12) from the rows of (6.14).

Let $k = rank(K)$. Since B^{-1} is positive definite, there exists a real number $\gamma > 0$ such that $\boldsymbol{v}^T B^{-1}\boldsymbol{v} \geq \gamma \boldsymbol{v}^T\boldsymbol{v}$ for all vectors $\boldsymbol{v}$. Let U be a matrix such that KU has orthonormal columns. Then

$$\begin{aligned}\boldsymbol{v}^T U^T K^T B^{-1} K U \boldsymbol{v} &= \boldsymbol{v}^T U^T K B^{-1} K U \boldsymbol{v} \\ &\geq \gamma \boldsymbol{v}^T U^T K K U \boldsymbol{v} \\ &= \gamma \boldsymbol{v}^T \boldsymbol{v},\end{aligned}$$

for any $\boldsymbol{v}$. Hence $X = U^T K B^{-1} K U$ is positive definite. Since X is a $k \times k$ matrix we have that the rank of X is k. However

$$k = rank(X) = rank(U^T K B^{-1} K U) \leq rank(K B^{-1} K) \leq k.$$

Hence $rank(KB^{-1}K) = k = rank(K)$ and the range of $KB^{-1}K$ is the entire range of K. Then there exists a matrix Y such that $YKB^{-1}K = K$. Left-multiplying (6.14) by Y, we obtain

$$YKB^{-1}A_{2LM} = K(I - B(S+B)^{-1}) \tag{6.15}$$

We can now recover (6.12) by selecting suitable rows of (6.15).

Note that each row of R_Γ coincides with some row of an R_{Γ_i}. Then there is a matrix

V which selects the appropriate rows of

$$K = W \begin{bmatrix} R_{\Gamma_1} \\ \vdots \\ R_{\Gamma_m} \end{bmatrix} \begin{bmatrix} R_{\Gamma_1}^T \cdots R_{\Gamma_m}^T \end{bmatrix},$$

such that $VK = R_\Gamma[R_{\Gamma_1}^T \cdots R_{\Gamma_m}^T]$. For this matrix V, we have

$$VYKB^{-1}A_{2LM} = R_\Gamma[R_{\Gamma_1}^T \cdots R_{\Gamma_m}^T](I - B(S+B)^{-1}),$$

which is the matrix on the left hand side of (6.12).

We have recovered the matrix on the left hand side of (6.12) and can now recover (6.11) via the relation (6.6) $= C_1$(6.11) $+ C_2$(6.12), as required. □

In the presence of cross points we can think of the 2LM method as a generalisation of OSM since we do not have a proof, like that of Lemma 3.2.1, showing equivalence between the methods. In fact it was shown, [25], that discretising a continuous OSM with cross points can lead to a problem that may stagnate or diverge, whereas the 2LM method deals with cross points systematically with no problems. However work has been carried out, [24], which shows that with a careful choice of Robin parameter at the cross points the convergence of the discrete OSM can be restored.

6.2 Preconditioners for the 2LM system

A goal of any domain decomposition method is to be *scalable* in the sense that as we increase the number of subdomains in the problem, the rate of convergence does not deteriorate too much. Most domain decomposition methods work by exchanging information about the solution between adjoining subdomains. As the problem is scaled up and more subdomains are involved this exchange of information locally can slow convergence as each subdomain

has to communicate with its neighbours. To achieve scalability a *coarse space correction* can be introduced. These corrections usually involve a projection onto a coarse space of the global problem allowing information to be exchanged globally between the subdomains and correct the local solutions.

For homogeneous problems and heterogeneous problems like ours, where the diffusion coefficient is constant within each subdomain and jumps only occur across the interface, coarse space corrections for classical Schwarz methods are well studied and there exists a rigorous convergence analysis, [50, 63]. For the OSM a coarse space correction was established in the case that the global domain of the problem is a cylinder and the subdomains are vertical strips, i.e. when no cross points are present, [12]. Coarse space corrections in the form of preconditioners have been developed for the 2LM method, with cross points, in both the case of homogeneous and heterogeneous problems, [39] and [47] respectively. In both papers a projection is used to define a preconditioner P and the preconditioned 2LM system:

$$P^{-1}A_{2LM} = P^{-1}\boldsymbol{\lambda},$$

is solved with the GMRES method. We follow the formulation of the preconditioners from the aforementioned papers and test their effectiveness numerically.

To construct our preconditoners we must first define a suitable projection onto a coarse space of our 2LM system. In a decomposition of a domain Ω into subdomains with cross points, *floating subdomains* will occur. A subdomain is said to "float" if no part of it intersects the natural boundary $\partial\Omega$. For each floating subdomain its associated Schur complement matrix, S_i, will have exactly one eigenvalue equal to 0. Then, recalling that $S = diag\{\alpha_1 S_1, \ldots, \alpha_m S_m\}$, the kernel of S, $ker(S) = \{\boldsymbol{v} \in \mathbb{R}^n : S\boldsymbol{v} = 0\}$, defines a coarse space of the 2LM method system. In particular the coarse space consists of piecewise constant functions with one degree of freedom per floating subdomain. Let E

denote the orthogonal projection onto $ker(S)$ then our first preconditioner is given by

$$P_1 = I - EBKB^{-1}E. \tag{6.16}$$

The above was introduced for a homogeneous problem however we note that (6.16) differs from the preconditioner presented in [39], which is given as $\tilde{P}_1 = I - EKE$. This occurs because in that paper only the case of one-sided Robin parameters are considered, i.e. $B = aI$ where a is the optimised Robin parameter. When we have one-sided parameters B will commute with K and (6.16) will simplify to $\tilde{P}_1$. If we allow for the Robin parameters to vary along the interface, such as when we have scaled one-sided parameters, B does not commute with K and the preconditioner is of the form shown in (6.16).

To see how we have arrived at this choice of preconditioner in (6.16) we start by claiming that the spectrum of $Q = B(S + B)^{-1}$ is contained in the interval $(0, 1]$, where the eigenvalues equal to 1 correspond to the floating subdomains in our decomposition. Using the spectral invariance property, that for any matrices C and D, $\sigma(C) = \sigma(D^{-1/2}CD^{1/2})$, we have that

$$\begin{aligned}
\sigma(Q) &= \sigma(B^{-1/2}QB^{1/2}) \\
&= \sigma(B^{1/2}(S + B)^{-1}B^{1/2}) \\
&= \sigma([B^{-1/2}(S + B)B^{-1/2}]^{-1}) \\
&= \sigma([B^{-1/2}SB^{-1/2} + I]^{-1}).
\end{aligned}$$

Then the spectrum of Q is of the form

$$\sigma(Q) = \left\{ \frac{1}{\lambda + 1} : \lambda \in \sigma(B^{-1/2}SB^{-1/2}) \right\}.$$

Now using the fact that S is symmetric positive semi-definite and B is symmetric

positive definite we have, for all $\boldsymbol{u} \in \mathbb{R}^n$ such that $\boldsymbol{u}^T\boldsymbol{u} = 1$, $\lambda = \boldsymbol{u}^T B^{-1/2} S B^{-1/2}\boldsymbol{u} = \boldsymbol{v}^T S\boldsymbol{v} \geq 0$, where $\boldsymbol{v} = B^{-1/2}\boldsymbol{u}$. In particular $\lambda = 0$ and hence $1 \in \sigma(Q)$ follows from the fact that

$$\begin{aligned} ker(B^{-1/2}SB^{-1/2}) &= \{\boldsymbol{v} \in \mathbb{R}^n : B^{-1/2}SB^{-1/2}\boldsymbol{v} = 0\} \\ &= \{\boldsymbol{v} \in \mathbb{R}^n : SB^{-1/2}\boldsymbol{v} = 0\} \\ &= \{\boldsymbol{v} = B^{1/2}\boldsymbol{w} \in \mathbb{R}^n : S\boldsymbol{v} = 0\} \\ &= B^{1/2}ker(S). \end{aligned}$$

Then the eigenvalues of Q equal to 1 correspond to the eigenvalues of S equal to 0 and hence those subdomains that float.

Knowing that $\sigma(Q) \subset (0,1) \cup \{1\}$ we can choose an orthonormal basis such that the eigenvectors of Q associated with the eigenvalues in the interval $(0,1)$ are listed first and those eigenvectors associated with eigenvalues equal to 1 second. Using this change of basis we can write in block form:

$$Q = \begin{bmatrix} Q_0 & 0 \\ 0 & I \end{bmatrix}, \qquad B = \begin{bmatrix} \tilde{B}_1 & 0 \\ 0 & \tilde{B}_2 \end{bmatrix}, \qquad K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}, \qquad E = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}.$$

Now, from (6.7), the 2LM system matrix is given by

$$A_{2LM} = \begin{bmatrix} (I - \tilde{B}_1 K_{11} \tilde{B}_1^{-1} - K_{11})Q_0 + K_{11} & -\tilde{B}_1 K_{12} \tilde{B}_2^{-1} \\ (-\tilde{B}_2 K_{21} \tilde{B}_1^{-1} - K_{21})Q_0 + K_{21} & I - \tilde{B}_2 K_{22} \tilde{B}_2^{-1} \end{bmatrix} \tag{6.17}$$

and our preconditioner (6.16) is of the form

$$P_1 = \begin{bmatrix} I & 0 \\ 0 & I - \tilde{B}_2 K_{22} \tilde{B}_2^{-1} \end{bmatrix}.$$

Then we see that preconditioner P_1 is obtained by replacing the top left block of (6.17) by the identity matrix and zeroing out the bottom left and top right blocks.

To use (6.16) in the preconditioned system $P_1^{-1}A_{2LM} = P_1^{-1}\boldsymbol{\lambda}$ we must form the inverse of P_1. This can be implemented in an efficient way as follows. Let $\mathbf{1}_{n_i}$ be the column vector of ones of length n_i where n_i is the size of the local stiffness matrix A_{N_i}. As $\mathbf{1}_{n_i}$ spans the kernel of A_{N_i} floating subdomains can be detected by checking if $A_{N_i}\mathbf{1}_{n_i} = 0$. Then let

$$\delta_i = \begin{cases} 1 \quad \text{if} \quad A_{N_i}\mathbf{1}_{n_{\Gamma_i}} = 0, \\ \\ 0 \quad \text{if} \quad A_{N_i}\mathbf{1}_{n_{\Gamma_i}} \neq 0 \end{cases}$$

and

$$\tilde{J} = blkdiag\left(\frac{\delta_1}{\sqrt{n_{\Gamma_1}}}\mathbf{1}_{n_{\Gamma_1}}, \ldots, \frac{\delta_p}{\sqrt{n_{\Gamma_p}}}\mathbf{1}_{n_{\Gamma_p}}\right),$$

where n_{Γ_i} is the number of vertices on $\partial\Omega_i \cap \Gamma$.

Now we form the matrix J by deleting the columns of $\tilde{J}$ that have entries all equal to zero. We have that $E = JJ^T$ where in our orthonormal basis

$$J = \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

and the inverse of P_1 is given by

$$\begin{aligned} P_1^{-1} &= \begin{bmatrix} I & 0 \\ 0 & (I - \tilde{B}_2 K_{22} \tilde{B}_2^{-1})^{-1} \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & (I - J^T B K B^{-1} J)^{-1} \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}}_{I-E} + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & (I - J^T B K B^{-1} J)^{-1} \end{bmatrix}}_{J(I-J^T B K B^{-1} J)^{-1} J^T}, \end{aligned}$$

i.e.

$$P_1^{-1} = I - JJ^T + J(I - J^T B K B^{-1} J)^{-1} J^T.$$

A second preconditioner for the 2LM system was introduced in [47] for heterogeneous problems. In that paper the heterogeneous problems considered are *multiscale*, allowing for the diffusion coefficient to vary greatly within the subdomains, as opposed to our problem where the coefficients are constant inside the subdomains. To take into account the multiscale coefficients, the authors of that paper use the coarse space defined by the piecewise constant functions of the kernel of S as well as functions that are "almost" in the kernel of S. These functions are found by solving a generalised eigenvalue problem. The reason this choice is made is that if diffusion is highly heterogeneous within a subdomain, and not just along the interface, the Schur complement matrix S_i for subdomain i will acquire isolated near zero eigenvalues. If the diffusion coefficients are constant within subdomains these near zero eigenvalues do not occur. As a result, in our case we are satisfied with the coarse space defined by the kernel of S.

Using the change of basis the second preconditioner is given by

$$P_2 = I - E + A_{2LM} E = \begin{bmatrix} I & -\tilde{B}_1 K_{12} \tilde{B}_2^{-1} \\ 0 & I - \tilde{B}_2 K_{22} \tilde{B}_2^{-1} \end{bmatrix}$$

which we see is obtained from (6.17) by replacing the top left block by the identity and zeroing out the lower left block.

To see why P_2 may be a better choice of preconditioner than P_1 consider the action of P_2^{-1} on the 2LM system matrix. Simplifying the notation we see that A_{2LM} and P_2^{-1} in block form are given by

$$A_{2LM} = \begin{bmatrix} W & X \\ Y & Z \end{bmatrix} \qquad \text{and} \qquad P_2^{-1} = \begin{bmatrix} I & -XZ^{-1} \\ O & Z^{-1} \end{bmatrix},$$

then

$$P_2^{-1} A_{2LM} = \begin{bmatrix} W - XZ^{-1}Y & O \\ Z^{-1}Y & I \end{bmatrix}. \tag{6.18}$$

Recall from the asymptotic convergence factor given by (4.33), the GMRES method will converge faster if the eigenvalues of the system matrix are clustered together away from the origin. Now since $P_2^{-1}A_{2LM}$ is block lower triangular we have that its spectrum is given by $\sigma(P_2^{-1}A_{2LM}) = \sigma(W - XZ^{-1}Y) \cup \{1\}$. From (6.18) we see that if we choose P_2 as a preconditioner it is in our best interest that block Z is the "bad" part of A_{2LM} since it is replaced by a block whose eigenvalues are equal to 1. Then P_2 will be an effective precondtioner if the eigenvalues of block $W - XZ^{-1}Y$ are clustered together close to the point $(1, 0)$ in $\mathbb{C}$. However just looking at the form of this block it is not clear if its eigenvalues are "good" in the sense that they are clustered away from the origin. To get this good clustering it may be necessary to use a different coarse space.

A similar block calculation for preconditioner P_1 gives us

$$P_1^{-1} A_{2LM} = \begin{bmatrix} W & X \\ Z^{-1}Y & I \end{bmatrix},$$

which due to its full block structure we cannot, at first glance, discern anything about its spectrum.

Like the first the inverse of this second preconditioner can be implemented in an efficient way:

$$P_2^{-1} = \begin{bmatrix} I & (\tilde{B}_1 K_{12} \tilde{B}_2^{-1})(I - \tilde{B}_2 K_{22} \tilde{B}_2^{-1})^{-1} \\ 0 & (I - \tilde{B}_2 K_{22} \tilde{B}_2^{-1})^{-1} \end{bmatrix}$$
$$= \underbrace{\begin{bmatrix} I & \tilde{B}_1 K_{12} \tilde{B}_2^{-1} \\ 0 & I \end{bmatrix}}_{I-(I-E)A_{2LM}E} \underbrace{\begin{bmatrix} I & 0 \\ 0 & (I - \tilde{B}_2 K_{22} \tilde{B}_2^{-1})^{-1} \end{bmatrix}}_{I-E+J(J^T A_{2LM} J)^{-1} J^T}.$$

giving us

$$P_2^{-1} = (I - (I - JJ^T)A_{2LM}JJ^T)(I - JJ^T + J(J^T A_{2LM} J)^{-1} J^T).$$

The size of $J^T A_{2LM} J$ is the same size as the coarse space which is much smaller than that of the 2LM system, so implementing P_2^{-1} will not be too costly.

To perform a similar estimate of the convergence of the 2LM method with GMRES for multiple subdomains as we did for the two subdomain case we would have to estimate the field of values $W(P_1^{-1} A_{2LM})$ and $W(P_2^{-1} A_{2LM})$. Within the period of this PhD we were unable to do so and a different bounding set of the spectrum such as the resolvent norm may need to be considered.

To see why the analysis we performed in Chapter 5, for two subdomain cases, cannot be used for the case of multiple subdomains consider the simplest case of three subdomains meeting at one cross point. The 2LM matrix for this problem will have the block structure:

$$A_{2LM} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{1x} \\ A_{21} & A_{22} & A_{23} & A_{2x} \\ A_{31} & A_{32} & A_{33} & A_{3x} \\ A_{x1} & A_{x2} & A_{x3} & A_{xx} \end{bmatrix},$$

where the subscript x's correspond to the crosspoint.

In the two subdomain case we had a 2×2 block structure, as shown in Corollary 5.1.2, that was tractable and we were able to estimate the field of values. The three subdomain case leads to a 4×4 block matrix which has eluded a similar analysis.

Though we cannot present analysis of the convergence of the 2LM method for heterogeneous problems with many subdomains we next perform some numerical experiments that test the effectiveness of our optimised Robin parameters and our choice of preconditioners.

6.3 Numerical experiments

We consider the model problem (1.2) with $f = 1$. A FEM discretisation is performed using piecewise linear triangular elements with mesh parameter h. We consider examples with $m > 2$ subdomains each having one of two diffusion coefficients α_1 or α_2 and Robin parameter p_1 or p_2. We use the optimised one-sided parameters, $p_1 = p_2 = \sqrt{\alpha_1 \alpha_2 s_{\min} s_{\max}}$ and scaled one-sided parameters $p_1 = \alpha_2 \sqrt{s_{\min} s_{\max}}$ and $p_2 = \alpha_1 \sqrt{s_{\min} s_{\max}}$, that were derived for the two subdomain case in Chapter 5. Now that we may have floating subdomains matrix S has eigenvalues equal to zero, then we take $s_{\min}$ to be the smallest non-zero eigenvalue of $\{S_i\}_{i=1}^m$ and $s_{\max}$ the largest eigenvalue.

Example 1

In the first example the domain is the unit square decomposed into a uniform grid of non-overlapping squares with sides of length H, distributed in such a way that they form a chequerboard pattern as shown in Figure 6.3. This results in subdomains with the same coefficients only meeting at cross points. Tables 6.1-6.6 show the number of iterations needed to reach a tolerance of 10^{-8} in the relative residual for different choices of h, $\omega = \alpha_2/\alpha_1$, and H. Results are calculated when there is no preconditioner and when the two preconditioners P_1 and P_2 are used.

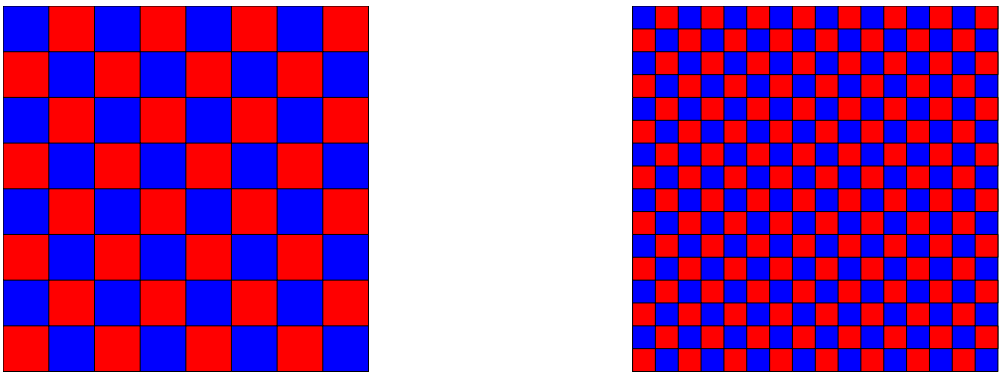

Figure 6.3: decomposition of unit square into a uniform grid of squares of length H, $H = 1/8$ (64 subdomains) on the left, $H = 1/16$ (256 subdomains) on right, subdomains with diffusion coefficient α_1 in blue, α_2 in red

	$h = 1/32$			$h = 1/64$			$h = 1/128$			$h = 1/256$		
Preconditioner:	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2
$\omega = 10^{-2}$	49	47	38	56	51	41	60	56	46	68	59	50
$\omega = 10^{-3}$	49	54	44	55	59	50	58	66	54	64	74	58
$\omega = 10^{-4}$	47	53	44	51	58	48	52	66	53	56	72	57
$\omega = 10^{-5}$	51	53	43	47	60	47	51	63	49	52	64	50

Table 6.1: Example 1: number of GMRES iterations when using one-sided Robin parameters, $H = 1/8$ (64 subdomains)

	$h = 1/32$			$h = 1/64$			$h = 1/128$			$h = 1/256$		
Preconditioner:	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2
$\omega = 10^{-2}$	40	36	28	40	37	30	42	35	30	46	38	30
$\omega = 10^{-3}$	33	27	21	36	27	22	36	28	23	37	29	24
$\omega = 10^{-4}$	26	21	14	28	20	15	29	20	16	30	22	16
$\omega = 10^{-5}$	18	20	14	17	19	15	18	19	14	18	21	15

Table 6.2: Example 1: number of GMRES iterations when using scaled one-sided Robin parameters, $H = 1/8$ (64 subdomains)

	$h = 1/32$			$h = 1/64$			$h = 1/128$			$h = 1/256$		
Preconditioner:	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2
$\omega = 10^{-2}$	77	45	35	86	51	39	98	57	43	108	65	47
$\omega = 10^{-3}$	99	61	47	111	70	53	128	78	55	145	89	65
$\omega = 10^{-4}$	105	75	54	119	87	63	133	90	64	142	101	72
$\omega = 10^{-5}$	105	84	60	115	86	62	125	97	67	136	108	74

Table 6.3: Example 1: number of GMRES iterations when using one-sided Robin parameters, $H = 1/16$ (256 subdomains)

	$h = 1/32$			$h = 1/64$			$h = 1/128$			$h = 1/256$		
Preconditioner:	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2
$\omega = 10^{-2}$	82	47	31	82	42	27	86	41	28	87	39	30
$\omega = 10^{-3}$	81	38	25	83	38	24	85	36	25	87	37	25
$\omega = 10^{-4}$	64	27	18	70	27	18	74	26	18	78	26	18
$\omega = 10^{-5}$	41	27	18	45	26	17	49	26	18	50	26	18

Table 6.4: Example 1: number of GMRES iterations when using scaled one-sided Robin parameters, $H = 1/16$ (256 subdomains)

	$h = 1/64$			$h = 1/128$			$h = 1/256$		
Preconditioner:	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2
$\omega = 10^{-2}$	107	41	30	132	46	34	157	55	39
$\omega = 10^{-3}$	167	58	44	184	66	48	209	77	55
$\omega = 10^{-4}$	196	66	51	222	79	56	248	92	65
$\omega = 10^{-5}$	217	72	51	234	87	63	248	88	60

Table 6.5: Example 1: number of GMRES iterations when using one-sided Robin parameters, $H = 1/32$ (1024 subdomains)

	$h = 1/64$			$h = 1/128$			$h = 1/256$		
Preconditioner:	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2
$\omega = 10^{-2}$	149	42	27	149	38	24	149	34	24
$\omega = 10^{-3}$	158	29	19	173	28	18	182	28	19
$\omega = 10^{-4}$	135	27	17	147	27	17	157	26	17
$\omega = 10^{-5}$	95	27	17	104	27	17	117	26	17

Table 6.6: Example 1: number of GMRES iterations when using scaled one-sided Robin parameters, $H = 1/32$ (1024 subdomains)

We observe for $H = 1/8$ that neither preconditioner performs that much better than when no preconditioner is used. This may be due to the relatively small number of subdomains used, 64 in this case. However when more subdomains are present, 256 and 1024 for $H = 1/16$ and $H = 1/32$ respectively, the preconditioners cut the iteration count markedly, with P_2 performing substantially better.

In terms of our optimised Robin parameters for the one-sided choice we seem to have lost the behaviour we observed in the two subdomain case, where increasing the jump in coefficients led to better convergence. On the other hand the scaled one-sided parameters retain this behaviour and as in the two subdomain case perform significantly better than the non-scaled parameters.

Example 2

In the first example the subdomains were "well mixed" in the sense that subdomains with the same diffusion coefficient only meet at cross points. We again consider an example where the unit square is decomposed into a uniform grid of squares of length H. Now the distribution of the diffusion coefficients creates the 2×2 chequerboard pattern as shown

in Figure 6.4. Whereas in the first example changing the value of H changed the pattern of the diffusion coefficients resulting in a different problem, in this example we keep the pattern and hence the problem fixed as we vary H. This results in subdomains with the same diffusion coefficients meeting at edges as well as cross points.

We must take care in our choice of Robin parameter at these edges and cross points as there is no jump in coefficients and locally we have a homogeneous problem between subdomains. In the case of one-sided Robin parameters no problem arises as we use the same Robin parameter on each side of the interface. However in the case of scaled one-sided parameters we make sure that subdomains that share the same diffusion coefficient have the appropriate Robin parameter.

For example on subdomains with diffusion coefficient α_1 the nodes on its interface Γ_i that meet any node from a subdomain with coefficient α_2 have scaled one-sided Robin parameter $p_1 = \alpha_2\sqrt{s_{\min}s_{\max}}$. While nodes on the interface that meet edges or cross points belonging only to subdomains with coefficient α_1 are given the Robin parameter $p_2 = \alpha_1\sqrt{s_{\min}s_{\max}}$.

Tables 6.7-6.12 show the number of iterations needed to reach a tolerance of 10^{-8} in the relative residual for different choices of h, $\omega = \alpha_2/\alpha_1$, and H. Results are calculated when there is no preconditioner and when the two preconditioners P_1 and P_2 are used.

In this example with less "well mixed" subdomains we observe that iteration counts are increased for all choices of parameter and precondtioner. Also we have lost the behaviour of improving convergence as the jump in coefficients increases. However, especially in the case of scaled parameters, the increase in the number of iterations as the jump increases is not too major. Again preconditioner P_2 with scaled one-sided parameters performs the strongest.

Where preconditioner P_1 fails is when we use scaled one-sided parameters and the jump in coefficients is large, as shown by the starred entries in the tables. For these entries the

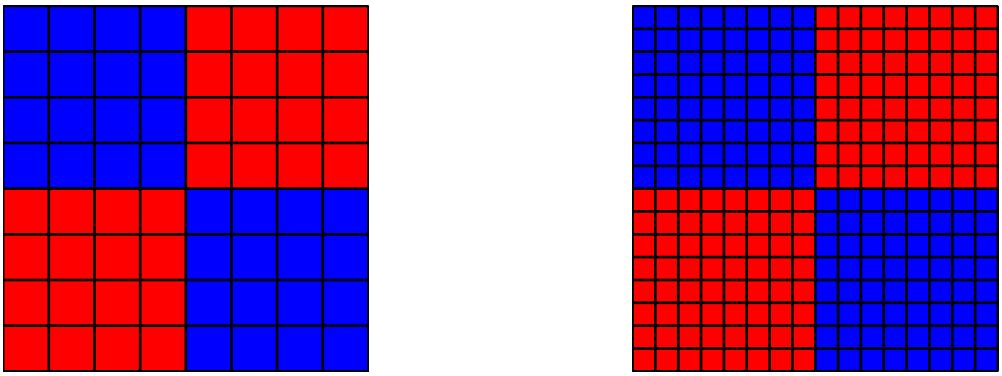

Figure 6.4: decomposition of unit square into a uniform grid of squares of length H, $H = 1/8$ (64 subdomains) on the left, $H = 1/16$ (256 subdomains) on right, subdomains with diffusion coefficient α_1 in blue, α_2 in red

	$h = 1/32$			$h = 1/64$			$h = 1/128$			$h = 1/256$		
Preconditioner:	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2
$\omega = 10^{-2}$	75	69	53	91	84	70	108	95	81	125	110	92
$\omega = 10^{-3}$	119	110	75	152	136	106	188	164	128	221	189	158
$\omega = 10^{-4}$	133	136	85	176	168	120	226	203	164	282	252	214
$\omega = 10^{-5}$	133	142	86	183	176	130	236	225	176	308	291	230

Table 6.7: Example 2: number of GMRES iterations when using one-sided Robin parameters, $H = 1/8$ (64 subdomains)

	$h = 1/32$			$h = 1/64$			$h = 1/128$			$h = 1/256$		
Preconditioner:	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2
$\omega = 10^{-2}$	42	50	39	43	57	43	50	65	49	57	64	57
$\omega = 10^{-3}$	59	49	46	49	56	48	51	67	57	54	81	68
$\omega = 10^{-4}$	47	46	44	55	53	54	50	62	64	57	74	58
$\omega = 10^{-5}$	51	20*	49	46	21*	45	53	24*	53	61	26*	64

Table 6.8: Example 2: number of GMRES iterations when using scaled one-sided Robin parameters, $H = 1/8$ (64 subdomains), starred entries did not converge to the correct solution

	$h = 1/32$			$h = 1/64$			$h = 1/128$			$h = 1/256$		
Preconditioner:	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2
$\omega = 10^{-2}$	117	80	43	138	93	53	164	106	67	188	119	77
$\omega = 10^{-3}$	195	122	60	255	151	79	314	184	101	370	219	128
$\omega = 10^{-4}$	252	167	67	326	210	93	403	269	119	477	339	166
$\omega = 10^{-5}$	280	203	71	358	263	93	432	347	132	520	433	188

Table 6.9: Example 2: number of GMRES iterations when using one-sided Robin parameters, $H = 1/16$ (256 subdomains)

	$h = 1/32$			$h = 1/64$			$h = 1/128$			$h = 1/256$		
Preconditioner:	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2
$\omega = 10^{-2}$	65	91	42	62	93	45	63	105	48	72	116	58
$\omega = 10^{-3}$	77	107	51	71	106	46	70	121	55	79	147	68
$\omega = 10^{-4}$	80	94	54	78	87	54	78	88	65	86	101	81
$\omega = 10^{-5}$	84	41*	50	74	38*	59	70	41*	52	75	40*	64

Table 6.10: Example 2: number of GMRES iterations when using scaled one-sided Robin parameters, $H = 1/16$ (256 subdomains), starred entries did not converge to the correct solution

	$h = 1/64$			$h = 1/128$			$h = 1/256$		
Preconditioner:	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2
$\omega = 10^{-2}$	283	96	41	339	107	52	388	122	62
$\omega = 10^{-3}$	522	156	55	649	188	74	771	288	99
$\omega = 10^{-4}$	681	219	64	869	277	87	1050	354	121
$\omega = 10^{-5}$	778	275	69	984	355	95	1200	469	133

Table 6.11: Example 2: number of GMRES iterations when using one-sided Robin parameters, $H = 1/32$ (1024 subdomains)

	$h = 1/64$			$h = 1/128$			$h = 1/256$		
Preconditioner:	I	P_1	P_2	I	P_1	P_2	I	P_1	P_2
$\omega = 10^{-2}$	95	164	63	100	173	62	116	194	70
$\omega = 10^{-3}$	127	212	72	118	234	73	127	258	76
$\omega = 10^{-4}$	132	85*	73	134	84*	71	135	89*	88
$\omega = 10^{-5}$	137	63*	79	127	64*	81	144	66*	72

Table 6.12: Example 2: number of GMRES iterations when using scaled one-sided Robin parameters, $H = 1/32$ (1024 subdomains), starred entries did not converge to the correct solution

GMRES algorithm stops once a tolerance of 10^{-8} in the relative residual has been met but the resulting solution is incorrect. This happens due to the relationship between the relative residual and the actual error of the solution.

To see this, assume the preconditioned 2LM method system $P^{-1}A_{2LM}\boldsymbol{\lambda} = P^{-1}\boldsymbol{c}$ is solved using GMRES with initial guess $\boldsymbol{\lambda}^0 = \mathbf{0}$, the zero vector. The GMRES method stops at step k if a set tolerance is reached in the relative residual:

$$\frac{\|P^{-1}\boldsymbol{c} - P^{-1}A_{2LM}\boldsymbol{\lambda}^k\|_2}{\|P^{-1}\boldsymbol{c} - P^{-1}A_{2LM}\boldsymbol{\lambda}^0\|_2} = \frac{\|\boldsymbol{r}^k\|_2}{\|P^{-1}\boldsymbol{c}\|_2}.$$

Let $\boldsymbol{e}^k = \boldsymbol{\lambda} - \boldsymbol{\lambda}^k$ be the actual error at step k. Then we have that

$$\begin{aligned}\|\boldsymbol{e}^k\|_2 = \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^k\|_2 = \|(P^{-1}A_{2LM})^{-1}(P^{-1}\boldsymbol{c}) - \boldsymbol{\lambda}^k\|_2 = \|(P^{-1}A_{2LM})^{-1}\boldsymbol{r}^k\|_2 \\ \leq \left\|\left|(P^{-1}A_{2LM})^{-1}\right|\right\|_2 \|\boldsymbol{r}^k\|.\end{aligned} \tag{6.19}$$

We also have that

$$\|P^{-1}\boldsymbol{c}\|_2 = \|P^{-1}A_{2LM}\boldsymbol{\lambda}\|_2 \leq \left|\left|\left|P^{-1}A_{2LM}\right|\right|\right|_2 \|\boldsymbol{\lambda}\|_2$$

and so

$$\frac{1}{\|\boldsymbol{\lambda}\|_2} \leq \frac{\left|\left|\left|P^{-1}A_{2LM}\right|\right|\right|_2}{\|P^{-1}\boldsymbol{c}\|_2}. \tag{6.20}$$

Combining (6.19) and (6.20) the relative error has the following bound:

$$\frac{\|\boldsymbol{e}^k\|_2}{\|\boldsymbol{\lambda}\|} \leq \underbrace{\left|\left|\left|P^{-1}A_{2LM}\right|\right|\right|_2 \left|\left|\left|(P^{-1}A_{2LM})^{-1}\right|\right|\right|_2}_{\kappa} \frac{\|\boldsymbol{r}^k\|_2}{\|P^{-1}\boldsymbol{c}\|_2}, \tag{6.21}$$

where κ is the 2-norm condition number of the preconditioned matrix $P^{-1}A_{2LM}$. If the condition number κ is large the GMRES method can stop after reaching a tolerance in the relative residual but produce a solution that has a large error.

We can further demonstrate this phenomenon of GMRES converging to the wrong solution by considering a small scale example. We want to solve $A\boldsymbol{x} = \boldsymbol{b}$ where

$$A = \begin{bmatrix} \epsilon & 1 \\ 0 & \epsilon \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{b} = \begin{bmatrix} \epsilon + 1 \\ \epsilon \end{bmatrix},$$

with $\epsilon > 0$ small. The true solution is $\boldsymbol{x} = [1, 1]^T$ and MATLAB has no problem solving this with the backslash command given a small $\epsilon = 10^{-8}$.

Now consider when we wish to solve the problem in MATLAB using the built in GMRES solver. Say we are given an initial guess to the solution $\boldsymbol{x}_0 = [1, 0]^T$. Although this vector is far from the true solution the residual $\boldsymbol{r}_0 = \boldsymbol{b} - A\boldsymbol{x}_0 = [10^{-8}, 0]^T$ is very small. In MATLAB, using this initial guess, GMRES converges in one iteration to a solution $\boldsymbol{x}_1$ that is even worse than $\boldsymbol{x}_0$ with a relative error of $\frac{\|\boldsymbol{x}-\boldsymbol{x_1}\|_2}{\|\boldsymbol{x}\|_2} = 3.5 \times 10^7$. Despite the huge error,

the residual, $\boldsymbol{r}_1 = [0, 0.5 \times 10^{-8}]^T$, is very small. The solution $\boldsymbol{x}_1$ given by GMRES is wrong by 7 orders of magnitude and the relative error is one billion percent. This problem arises due to how poorly conditioned the matrix A is.

Note that even using a precondtioner P can lead to the same problem if $P^{-1}A$ is poorly conditioned. Using the Jacobi preconditioner with GMRES in MATLAB on the problem we get the same large error but small residual.

Both the current example and the previous one demonstrate that with the choice of preconditioner P_2 and scaled one-sided parameters for the 2LM method we have an efficient scheme to solving large scale heterogeneous problems with many subdomains and cross points. Also that the more "well mixed" the problem is the faster convergence should be.

Example 3

We finish with an example that shows how we can use the 2LM method to solve a heterogeneous problem that arises from a real world application, the problem of seepage under a dam. When building a dam the engineers must take into account the material that the dam will sit on. The porosity of the ground underneath will determine how much water seeps under the dam. If too much water flows underneath, the effectiveness of the dam may be compromised.

The domain Ω is shown in Figure 6.5 and, although the geometry is not based on any real world dam, it is representative of dams shaped like that of the Fanshawe dam, located in Canada, shown in Figure 6.6. The problem we consider has been studied before for homogeneous media in [46], here we present results for a heterogeneous problem where under the dam there are two types of media of vastly different permeability. We are interested in how different configurations of the material affect seepage under the dam.

The governing equation is of a stationary state of groundwater flowing through porous

media:

$$\begin{cases} -\nabla \cdot (k(\boldsymbol{x})\nabla h) = 0 & \text{in} \quad \Omega \\ h = g_D & \text{on} \quad \partial\Omega_D \\ \frac{\partial h}{\partial n} = g_N & \text{on} \quad \partial\Omega_N, \end{cases} \tag{6.22}$$

where h is the total hydraulic head and $k(\boldsymbol{x})$ the hydraulic permeability coefficient. The domain $\Omega \subset \mathbb{R}^n$ is partitioned into non-overlapping subdomains that have either permeability coefficient $k(\boldsymbol{x}) = k_1 = 1$ or $k(\boldsymbol{x}) = k_2 = 10^{-5}$. Subdomains with coefficient k_1 correspond to media that is highly permeable like gravel or fractured rocks. Subdomains with coefficient k_2 correspond to media that is semi-impermeable such as clay or sandstone. The permeability coefficients for different materials can be found in [1].

The upstream and downstream water levels are assumed to be 10 and 1 respectively, which implies Dirichlet boundary conditions $h|_{D_1} = g_{D_1} = 10$ and $h|_{D_2} = g_{D_2} = 1$. Moreover we assume the dam is made of impermeable material and that there is no flow coming through the bottom, left or right sides of the domain, implying zero Neumann boundary conditions, $\partial h/\partial n = g_N = 0$ on those boundary components.

Figure 6.7 shows the solutions calculated using the 2LM method for various arrangements of subdomains. The plots on the left show the decomposition of the domain into the two types of media. Brown subdomains represent material with high permeability while the grey subdomains represent semi-impermeable material. The colour scale for the plots on the left, showing the solution for each configuration, is read as blue representing water seeping under the dam. As the colour changes from blue to red, less water has flowed through this material.

There are some physical conclusions we can observe from the solution for different configurations. Starting from the top solution in Figure 6.7, this represents a homogeneous

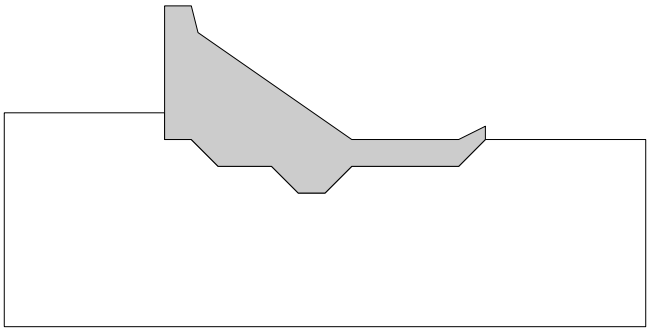

Figure 6.5: domain (in white) under a dam (in grey)

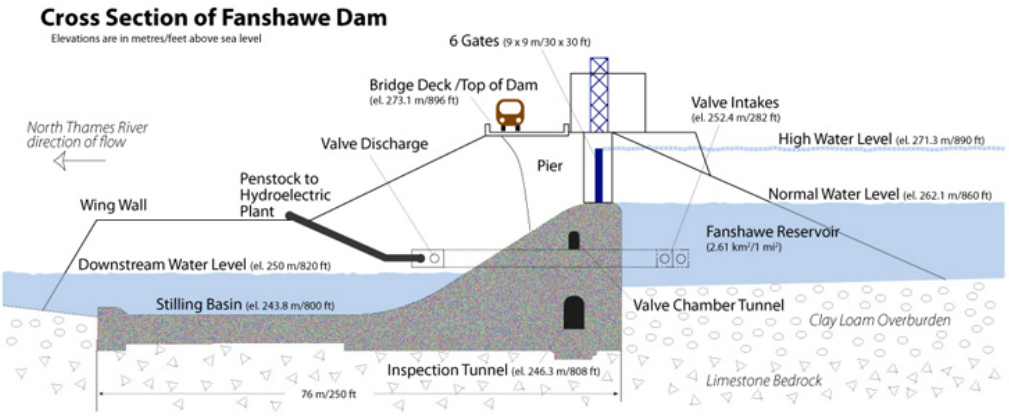

Figure 6.6: cross section of the Fanshawe dam, Ontario, Canada. Retrieved from http://thamesriver.on.ca/water-management/flood-control-structures/fanshawe-dam/

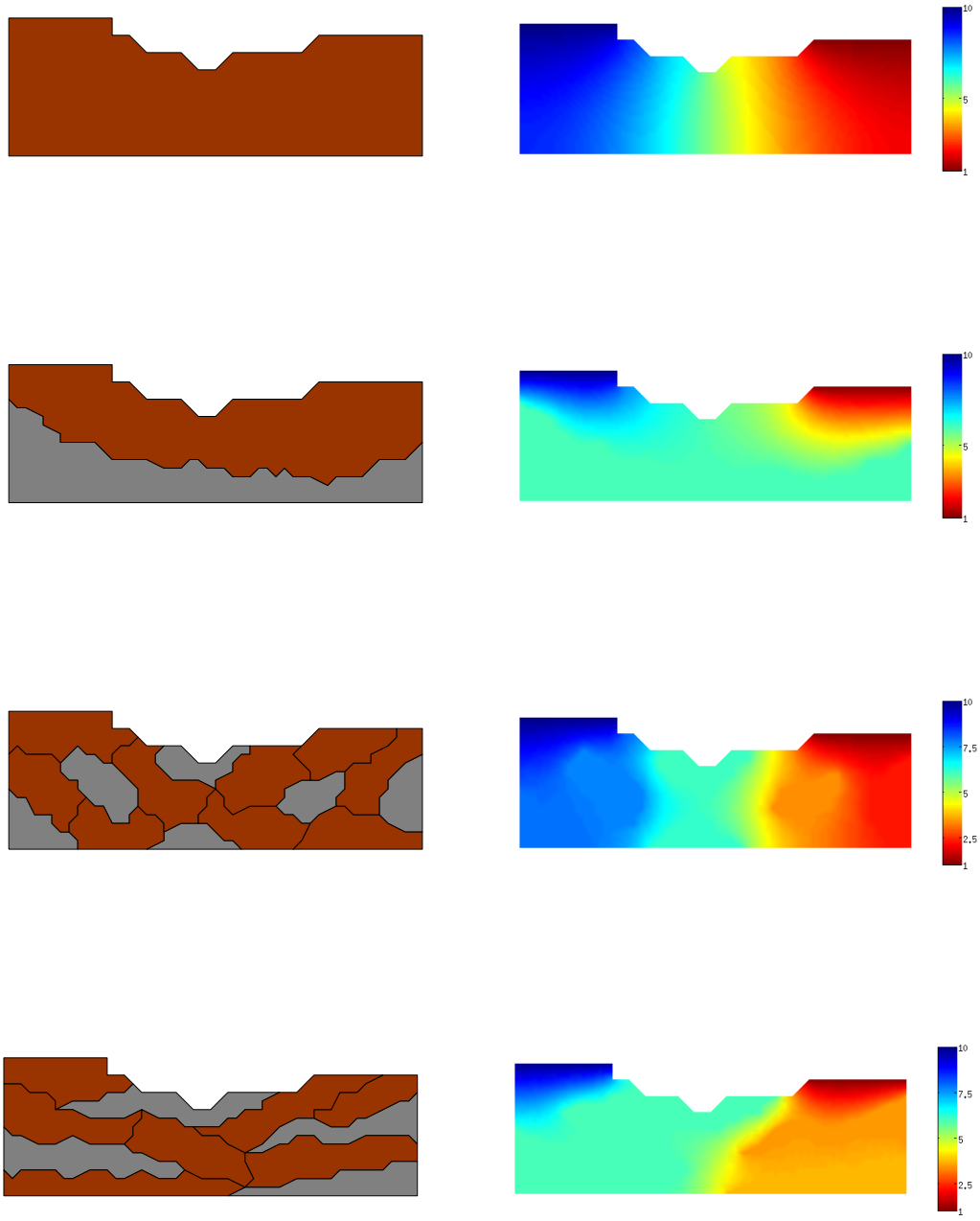

Figure 6.7: solutions (on the right) for the problem of seepage under a dam for various decompositions of the domain into subdomains with high permeability (in brown) and subdomains with low permeability (in grey)

case were the media under the dam is all of the permeable type. We see that this permeability allows very little water to seep under the dam as most of it flows into the material on the left hand side.

The next configuration has two subdomains representing a layer of semi-impermeable material underneath the permeable material. This layer of semi-impermeable material is allowing more water to flow, from left to right, through the permeable material and under the dam than in the homogeneous case.

The third and fourth configurations have multiple subdomains with the first representing the semi-impermeable material arranged in pockets and the other with the material arranged in layers. We see that both configurations allow more water to flow under the dam than the homogeneous case but less as compared with the two subdomain case. Of the two, the configuration with pockets performs better than the one with layers.

Chapter 7

Conclusion

The aim of this thesis was to study the OSM and the 2LM method for heterogeneous problems. For a model problem with two subdomains we have formulated the 2LM method and shown that we can use it to derive the global solution to a BVP (Lemma 3.1.1). When the 2LM system is solved with a Richardson iteration we have proven its equivalence to the OSM (Lemma 3.2.1).

To derive a convergence estimate for the 2LM method when solved using GMRES we have approximated the field of values of the system matrix by a rectangle (Corollary 5.1.2) and derived optimised one-sided and scaled-one-sided Robin parameters that speed up the convergence of the GMRES method (Theorem 5.2.1 and Theorem 5.2.4 respectively). For the case of one-sided parameters, in constructing a conformal map from the interior of the unit disc to exterior of the rectangle (Lemma 5.2.2) we have derived an estimated asymptotic convergence factor for the 2LM method with GMRES and shown the behaviour as the mesh is refined and the jump in coefficients becomes large (Theorem 5.2.3). Numerical experiments at the end of Chapter 5 confirm our theoretical results.

Finally, in Chapter 6, we formulated the 2LM method for heterogeneous problems with many subdomains and cross points (Lemma 6.1.1 and Theorem 6.1.2). We have constructed

preconditioners for the 2LM method system and tested their effectiveness numerically.

Future work that arises from the results of this thesis include the derivation of the optimised two-sided parameters for the 2LM method. If they could be found we would expect them to perform better than one-sided and scaled one-sided parameters, as was the case with the OSM.

Further work can be done to analyse the convergence of the 2LM method for heterogeneous problems with many subdomains. We have shown numerically that with a suitable choice of Robin parameter and preconditioner we can achieve favourable convergence in the case of many subdomains and cross points. The theory underpinning the numerical results can be derived using a similar estimate of GMRES convergence as we performed for the two subdomain case. As the field of values of the system matrix is likely to include the origin, the estimate can instead be calculated using the resolvent norm of the system matrix.

The analysis presented could be further extend to include the case when we have multiple subdomains with three or more different diffusion coefficients rather than the case of two presented here in the numerical experiments at the end of chapter 6. This would lead to studying the case were we allow jumps within the subdomains not just along the interface.

Bibliography

[1] J. Bear, *Dynamics of fluids in porous media*, Courier Corporation, 2013.

[2] B. Beckermann, S. A. Goreinov, and E. E. Tyrtyshnikov, *Some remarks on the Elman estimate for GMRES*, SIAM journal on Matrix Analysis and Applications, 27 (2005), pp. 772–778.

[3] S. C. Brenner, *The condition number of the Schur complement in domain decomposition*, Numer. Math., 83 (1999), pp. 187–203.

[4] R. Bulirsch and J. Stoer, *Introduction to numerical analysis*, Springer, 2002.

[5] M. Crouzeix, *Bounds for analytical functions of matrices*, Integral Equations and Operator Theory, 48 (2004), pp. 461–477.

[6] ——, *Numerical range and functional calculus in Hilbert space*, Journal of Functional Analysis, 244 (2007), pp. 668–690.

[7] Q. Deng, *Timely communicaton: An analysis for a nonoverlapping domain decomposition iterative procedure*, SIAM J. Sci. Comput., 18 (1997), pp. 1517–1525.

[8] T. A. Driscoll, K.-C. Toh, and L. N. Trefethen, *From potential theory to matrix iterations in six steps*, SIAM Rev., 40 (1998), pp. 547–578.

[9] T. A. Driscoll and L. N. Trefethen, *Schwarz-Christoffel Mapping*, Cambridge University Press, 2002.

[10] S. W. Drury and S. Loisel, *Sharp condition number estimates for the symmetric 2-Lagrange multiplier method*, in Domain Decomposition Methods in Science and Engineering XX, Springer, 2013, pp. 255–261.

[11] O. Dubois, *Optimized Schwarz methods for the advection-diffusion equation and for problems with discontinuous coefficients*, PhD thesis, McGill University, 2007.

[12] O. Dubois, M. J. Gander, S. Loisel, A. St-Cyr, and D. B. Szyld, *The optimized Schwarz method with a coarse grid correction*, SIAM Journal on Scientific Computing, 34 (2012), pp. A421–A458.

[13] O. Dubois and S. Lui, *Convergence estimates for an optimized Schwarz method for PDEs with discontinuous coefficients*, Numer. Algorithms, 51 (2009), pp. 115–131.

[14] M. Eiermann, *Semiiterative Verfahren für nichtsymmetrische lineare Gleichungssysteme*, PhD thesis, 1990.

[15] ——, *Fields of values and iterative methods*, Linear Algebra and its Applications, 180 (1993), pp. 167–197.

[16] H. C. Elman, *Iterative methods for large, sparse, nonsymmetric systems of linear equations*, PhD thesis, Yale University New Haven, Conn, 1982.

[17] M. Embree, *How descriptive are GMRES convergence bounds?*, tech. rep., Oxford University Computing Laboratory, 1999.

[18] V. Faber, W. Joubert, E. Knill, and T. Manteuffel, *Minimal residual method stronger than polynomial preconditioning*, SIAM Journal on Matrix Analysis and Applications, 17 (1996), pp. 707–729.

[19] C. Farhat, A. Macedo, M. Lesoinne, F.-X. Roux, F. Magoulés, and A. De La, Bourdonnaie, *Two-level domain decomposition methods with Lagrange multipliers for the fast iterative solution of acoustic scattering problems*, Comput. Methods in Appl. Mech. Eng., 184 (2000), pp. 213–239.

[20] M. J. Gander, *Optimized Schwarz methods*, SIAM J. Numer. Anal., 44 (2006), pp. 699–731.

[21] ——, *Schwarz methods over the course of time*, Electron. Trans. Numer. Anal., 31 (2008), pp. 228–255.

[22] M. J. Gander and O. Dubois, *Optimized Schwarz methods for a diffusion problem with discontinuous coefficient*, Numer. Algorithms, (2014). doi:10.1007/s11075-014-9884-2.

[23] M. J. Gander, L. Halpern, and F. Nataf, *Optimized Schwarz methods*, in Twelth International Conference on Domain Decomposition Methods, Bergen, 2001, Domain Decomposition Press, pp. 15–28.

[24] M. J. Gander and F. Kwok, *Best Robin parameters for optimized Schwarz methods at cross points*, SIAM J. Sci. Comput., 34 (2012), pp. A1849–A1879.

[25] ——, *On the applicability of Lions' energy estimates in the analysis of discrete optimized Schwarz methods with cross points*, in Domain Decomposition Methods in Science and Engineering XX, Springer, 2013, pp. 475–483.

[26] M. J. Gander, F. Magoules, and F. Nataf, *Optimized Schwarz methods without overlap for the Helmholtz equation*, SIAM J. Sci. Comput., 24 (2002), pp. 38–60.

[27] G. H. Golub and C. F. Van Loan, *Matrix computations*, John Hopkins University Press, 1996.

[28] A. Greenbaum, *Iterative methods for solving linear systems*, vol. 17, SIAM, 1997.

[29] A. Greenbaum and L. Gurvits, *Max-min properties of matrix factor norms*, SIAM Journal on Scientific Computing, 15 (1994), pp. 348–358.

[30] A. Greenbaum, V. Pták, and Z. e. k. Strakoš, *Any nonincreasing convergence curve is possible for GMRES*, SIAM journal on matrix analysis and applications, 17 (1996), pp. 465–469.

[31] A. Greenbaum and Z. Strakos, *Matrices that generate the same Krylov residual spaces*, Springer, 1994.

[32] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM Journal on Scientific Computing, 15 (1994), pp. 359–368.

[33] N. Greer and S. Loisel, *The optimised Schwarz method and the two-Lagrange multiplier method for heterogeneous problems in general domains with two general subdomains*, Numerical Algorithms, 69 (2015), pp. 737–762.

[34] E. Hille, *Analytic function theory*, vol. 2, American Mathematical Soc., 2005.

[35] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*, Cambridge University Press, 1991.

[36] ——, *Matrix analysis*, Cambridge university press, 2012.

[37] C. R. Johnson, *Numerical determination of the field of values of a general complex matrix*, SIAM J. Numer. Anal., 15 (1978), pp. 595–602.

[38] W. Joubert, *A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems*, SIAM Journal on Scientific Computing, 15 (1994), pp. 427–439.

[39] A. Karangelis and S. Loisel, *Condition number estimates and weak scaling for 2-level 2-Lagrange multiplier methods for general domains and cross points*, SIAM Journal on Scientific Computing, 37 (2015), pp. C247–C267.

[40] T. Kato, *Some mapping theorems for the numerical range*, Proceedings of the Japan Academy, 41 (1965), pp. 652–655.

[41] ——, *Perturbation theory for linear operators*, vol. 132, Springer Science & Business Media, 2013.

[42] J. Liesen and Z. Strakos, *Krylov subspace methods: principles and analysis*, Oxford University Press, 2012.

[43] J. Liesen and P. Tichỳ, *Convergence analysis of Krylov subspace methods*, GAMM-Mitteilungen, 27 (2004), pp. 153–173.

[44] P.-L. Lions, *On the Schwarz alternating method. iii: a variant for nonoverlapping subdomains*, in Third international symposium on domain decomposition methods for partial differential equations, vol. 6, SIAM, Philadelphia, PA, 1990, pp. 202–223.

[45] S. Loisel, *Condition number estimates for the nonoverlapping optimized Schwarz method and the 2-Lagrange multiplier method for general domains and cross points*, SIAM J. Numer. Anal., 51 (2013), pp. 3062–3083.

[46] S. Loisel and H. Nguyen, *An optimal schwarz preconditioner for a class of parallel adaptive finite elements*, Manuscript submitted for publication, (2016).

[47] S. Loisel, H. Nguyen, and R. Scheichl, *Optimized Schwarz and 2-Lagrange multiplier methods for multiscale elliptic PDEs*, SIAM Journal on Scientific Computing, 37 (2015), pp. A2896–A2923.

[48] S. Lui, *A Lions non-overlapping domain decomposition method for domains with an arbitrary interface*, IMA J. Numer. Anal., 29 (2009), pp. 332–349.

[49] Y. Maday and F. Magoulès, *Optimized Schwarz methods without overlap for highly heterogeneous media*, Comput. Methods in Appl. Mech. Eng., 196 (2007), pp. 1541–1553.

[50] T. P. A. Mathew, *Domain decomposition methods for the numerical solution of partial differential equations*, Springer, 2008.

[51] N. M. Nachtigal, S. C. Reddy, and L. N. Trefethen, *How fast are nonsymmetric matrix iterations?*, SIAM Journal on Matrix Analysis and Applications, 13 (1992), pp. 778–795.

[52] O. Nevanlinna, *Convergence of interations for linear equations*, Springer, 1993.

[53] C. Pearcy et al., *An elementary proof of the power inequality for the numerical radius.*, The Michigan Mathematical Journal, 13 (1966), pp. 289–291.

[54] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical mathematics*, Springer, 2000.

[55] A. Quarteroni and A. Valli, *Numerical approximation of partial differential equations*, vol. 23, Springer Science & Business Media, 2008.

[56] A. Quarteroni, A. Valli, et al., *Domain decomposition methods for partial differential equations*, Oxford University Press, 1999.

[57] F.-X. Roux, F. Magoulès, S. Salmon, and L. Series, *Optimization of interface operator based on algebraic approach*, in Fourteenth International Conference on Domain Decomposition Methods, Bergen, 2003, Domain Decomposition Press, pp. 297–304.

[58] Y. Saad, *Iterative methods for sparse linear systems*, SIAM, 2003.

[59] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput, 7 (1986), pp. 856–869.

[60] H. A. Schwarz, *Über einen grenzübergang durch altermierendes verfahren*, Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich, 15 (1870), pp. 272–286.

[61] K. C. Toh, *Matrix Approximation Problems and Nonsymmetric Iterative Methods*, PhD thesis, Cornell University, 1996.

[62] K.-C. Toh, *GMRES vs. ideal GMRES*, SIAM Journal on Matrix Analysis and Applications, 18 (1997), pp. 30–36.

[63] A. Toselli and O. Widlund, *Domain decomposition methods: algorithms and theory*, Springer, 2005.

[64] L. N. Trefethen, *Pseudospectra of matrices*, Numerical analysis, 91 (1991), pp. 234–266.

[65] L. N. Trefethen and D. Bau III, *Numerical linear algebra*, SIAM, 1997.

[66] O. B. Widlund, *Accomodating irregular subdomains in domain decomposition theory*, in Domain decomposition methods in science and engineering XVIII, Springer, 2009, pp. 87–98.